

THE PSYCHOLOGICAL CONTROL OF IMPLICIT BIASES: A DYNAMICAL SYSTEMS PERSPECTIVE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Michael Thomas Wojnowicz

May 2012

© 2012 Michael Thomas Wojnowicz

ALL RIGHTS RESERVED

THE PSYCHOLOGICAL CONTROL OF IMPLICIT BIASES: A DYNAMICAL SYSTEMS PERSPECTIVE

Michael Thomas Wojnowicz, Ph.D.

Cornell University 2012

Much of social psychological theorizing is entrenched in a dualism between two distinctive mental systems – one associative, the other rule-based. In particular, in the field of evaluations, the contemporary dual systems approach emphasizes the separated existence of distinct implicit and explicit attitudes. However, in recent times, theoreticians have been seeking an understanding of social psychological topics through models that can handle real-time interactivity between component parts. Thus, this dissertation applies the framework of dynamical systems towards key social psychological topics typically construed through dual systems theory. In Chapter 2, we provide evidence that explicit evaluations are gradually unfolding from the self-organization of multiple biases. By analyzing hand-movement trajectories in an explicit attitude report task, we show that while our participants are about equally likely to report liking white people and black people, their formations of these two responses show qualitatively distinct processing dynamics. These findings support the notion that the mind hosts a continuously evolving blend of evaluative decisions from which the eventual explicit decision emerges. In Chapter 3, we provide preliminary evidence that the dynamics of formulating an explicit evaluative judgment is even biased by subliminal evaluative conditioning. These findings would challenge the notion that

explicit and implicit attitudes partake in two distinct psychological "channels," suggesting instead that a dynamically interactive mind underlies the preparation of an eventual explicit decision. Finally, in Chapter 4, we sketch out a dynamical systems approach to motivated control. We provide dynamical systems interpretations for three constituent aspects of control – selection, goal pursuit, and top-down flexibility, and thereby craft a perspective on motivated control which respects the existence of specialized neurobiological systems, but creates space for more than two of them, and allows them to continually interact.

BIOGRAPHICAL SKETCH

Michael Wojnowicz was born on April 18, 1979 in Jacksonville, Florida. He studied at a Catholic high school, Bishop Kenny High School, in Jacksonville, Florida, and he then went on to study psychology at the University of Florida. Upon graduating, he spent the next three years teaching in Gainesville, New York City and Prague, Czech Republic. During these experiences in racially diverse classrooms, he became especially motivated by the question of how people propagate stereotyping and racism without knowing it. In 2004, Mike began doctoral study in psychology at Cornell University to pursue this question. Early on in these studies, he connected with his long-standing interest in dynamical systems, encouraged especially by his experience at the Santa Fe Institute's Complex Systems Summer School. His personal interests outside of graduate school include ultimate frisbee, tango, improvisational acting, and cross-country running.

This document is dedicated first of all to my committee members, without whom this dissertation would not be. Melissa, you have been the most terrific advisor I could possibly imagine. I am so very thankful for your encouragement and support – and for being open-minded and willing to go on this intellectual journey. Spivey, your mini-lecture on dynamical systems (in preparation for your NYU talk) completely changed me. I left thinking, “I would do anything to approach psychology like THAT.” Your encouragement led to Chapter 2 of this dissertation. Melissa, my perspective on automaticity has developed constantly through my many interactions with you. It is to you that I dedicate Chapter 3. Barb, your insistence that “the brain is not an SRN” during my A-exam meeting was a beautiful comment that is single-handedly responsible for Chapter 4 of this dissertation. Finally, David, your mathematical approach to psychology has been a major source of inspiration, including for my future studies of mathematics at the University of Washington. In addition to my committee, I would like to thank the many positive influences in my personal life – my family, my girlfriend, and all my friends. I love you all and thank you for many wonderful experiences. Finally, I would to give special thanks to those who have taught me the kinds of wisdom that cannot be taught in books – to my acting teacher here in Ithaca, Eliza Van Cort, for teaching me Meisner; to my father, for teaching me both compassion and logic; to my mother, who taught me love; and to my high-school cross-country coach, who taught me to never give up.

ACKNOWLEDGEMENTS

I would like to acknowledge the frequent financial support of the Cognitive Science Program at Cornell University, especially for full support in the summer of 2008. I would also like to most gratefully acknowledge the full year of support from the Dallenbach Fellowship during the 2007-2008 academic year. Finally, I would like to acknowledge the National Science Foundation for their support of my studies at the Santa Fe Institute in the summer of 2007.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Table of Contents	vi
List of Tables	ix
List of Figures	x
1 Introduction	1
1.1 The Automaticity Revolution in Social Psychology	1
1.1.1 The existence of implicit social information	1
1.1.2 The distinction between implicit and explicit cognition	5
1.2 Theoretical Perspectives on Automaticity	6
1.2.1 The dualistic perspective	6
1.2.2 The dynamical perspective	8
1.3 Empirical Goal: How do implicit biases influence the dynamics of forming an explicit judgment?	12
1.3.1 Chapter 2: The influence of aversive racism	17
1.3.2 Chapter 3: The influence of subliminal conditioning	18
1.4 Theoretical Goal: How does a dynamic mental system accomplish psychological control?	19
1.4.1 Chapter 4: A dynamical interactive model of psychological control	20
2 The Dynamical Influence of Aversive Racism On Explicit Evaluative Processing	22
2.1 Overview	22
2.2 Study 1	29
2.2.1 Study 1: Methods	29
2.2.2 Study 1: Results and Discussion	30
2.3 Study 2	37
2.3.1 Study 2: Methods	37
2.3.2 Study 2: Results and Discussion	38
2.4 Study 3	38
2.4.1 Study 3: Methods	39
2.4.2 Study 3: Results	39
2.5 Discussion	41

3	The Dynamical Influence of Subliminal Conditioning On Explicit Evaluative Processing	42
3.1	Overview	42
3.1.1	A "dual channels" model of evaluative decisions	42
3.2	Pilot Study	48
3.2.1	Pilot Study: Methods	48
3.2.2	Pilot Study: Data Analysis Methods	51
3.2.3	Pilot Study: Results and Discussion	54
3.3	Study 4	56
3.3.1	Study 4: Methods	56
3.3.2	Study 4: Results and Discussion	57
3.4	Study 5	59
3.4.1	Study 5: Methods	59
3.4.2	Study 5: Results and Discussion	59
3.5	Study 6	61
3.5.1	Study 6: Methods	61
3.5.2	Study 6: Results and Discussion	62
3.6	Conclusions	64
4	A Dynamical Systems Approach to Psychological Control	66
4.1	Introduction: The interactive control problem	66
4.1.1	Psychological control is required to override implicit biases	66
4.1.2	What is psychological control? The dualistic perspective	69
4.1.3	The three capacities of a psychological controller	71
4.1.4	What is psychological control? The need for an interactive theory	74
4.1.5	O'Reilly's four principles of real-time mental processing	75
4.1.6	The dynamical systems perspective on mental processing	77
4.1.7	Goal of this chapter	88
4.1.8	A note on computation vs. systems	91
4.2	Selection: Dynamic contraction	93
4.2.1	Selection via point attractors	94
4.2.2	Competitive selection via multistability	97
4.2.3	Thesis: Selection as dynamic contraction	100
4.3	Strategic goal pursuit: Attractor landscapes acknowledging distant end-states	102
4.3.1	Dualistic perspectives throughout the history of psychology	103
4.3.2	Toward an interactive perspective on goal pursuit	112

4.3.3	Brain implementation in multiple interacting systems	120
4.3.4	Thesis: Goal pursuit as dopamine-sculpted attractor landscapes	128
4.4	Flexible top-down processing: The flexible selection of attractor landscapes	133
4.4.1	Working memory: The information processing requirement	135
4.4.2	Prefrontal cortex: Neurophysiological properties	143
4.4.3	Control parameters: The dynamical systems concept	148
4.4.4	Thesis: Flexible top-down processing as the adaptive selection of attractor landscapes	153
4.5	Closing thoughts	158
4.5.1	Distributed interactive control	158
4.5.2	The "dynamic projections" hypothesis of psychological control	164
5	Conclusion	173
5.1	Summary of Dissertation	173
5.2	Limitations	174
	Bibliography	180

LIST OF TABLES

1.1	A normalized recurrence model of a dynamically self-organizing deliberative evaluation. A node's activation is represented by darkness.	16
4.1	Dopamine fires in response to mispredicted reward, from Fiorillo, Tobler, & Schultz(2003)	124
4.2	Flexible top-down control: A three-way correspondence between model, behavior, and implementation	154

LIST OF FIGURES

2.1	<i>Studies 1 and 2: Mean mouse-movement trajectories towards evaluative targets for the black people and white people stimuli</i>	31
2.2	<i>Study 1: Distributions of trajectory curvature show no evidence of bimodality</i>	32
2.3	<i>Study 1: Mean velocities towards the evaluative decision for the black people and white people conditions in normalized time . . .</i>	35
2.4	<i>Study 1: Hand trajectories exhibit greater spatial disorder during evaluations of black people than white people</i>	36
2.5	<i>Study 3: Mean mouse-movement trajectories towards evaluative targets for the stimuli African Americans and Caucasians in the third study</i>	40
3.1	<i>Primary findings of Rydell & McConnell 2006</i>	45
3.2	<i>A technique for measuring the spatial deflection of hand-movement trajectories</i>	52
3.3	<i>Pilot Experiment: Mean hand-movement trajectories towards the like decision for negatively conditioned and positively conditioned stimuli</i>	55
3.4	<i>Study 4: Mean hand-movement trajectories towards the like decision for negatively conditioned and positively conditioned stimuli</i>	58
3.5	<i>Study 5: Mean hand-movement trajectories towards the dislike decision for negatively conditioned and positively conditioned stimuli</i>	60
3.6	<i>Study 6: Mean hand-movement trajectories towards the like decision for negatively conditioned and positively conditioned stimuli</i>	63
4.1	<i>Dual systems of cognition, from Kahneman & Frederick 2002 . . .</i>	70
4.2	<i>An attractor landscape.</i>	82
4.3	<i>A point attractor in state space, from Chris Eliasmith</i>	94
4.4	<i>A bowl shaped attractor landscape, from Seung 1996</i>	96
4.5	<i>The Lotka-Volterra model for two competing species</i>	99
4.6	<i>Many recurrent loops within the basal ganglia, from Bar-Gad et al. 2003</i>	121
4.7	<i>The actor/critic model of the basal ganglia, from Niv et al. 2010 . .</i>	122
4.8	<i>A methodology for documenting state improvement, from Seymour et al. 2004</i>	126
4.9	<i>fMRI-BOLD results suggesting state improvement, from Seymour et al. 2004</i>	127

4.10	Seventy-two regions of cortex in connectivity space, from Young 1993.	146
4.11	A saddle-node bifurcation, from Strogatz 1994.	149
4.12	Control is distributed throughout interactive dynamical system. Figure adapted from Botvinick, Niv, & Barto 2009 to incorporate table	161
4.13	Hypothetical distributed representations for libraries, good, and parties	170
4.14	The relative positivity of libraries vs. parties depends upon how the prefrontal cortex projects high-dimensional representations onto low-dimensional subspaces	171

CHAPTER 1

INTRODUCTION

1.1 The Automaticity Revolution in Social Psychology

1.1.1 The existence of implicit social information

External Primes

A rich set of experiments during the "automaticity revolution" of experimental social psychology has demonstrated that the way people think, decide, and behave can be strongly influenced by environmental primes - that is, subtle exposures to information that an individual does not consciously notice or recognize. For instance, when participants unscrambled twenty sentences containing words related to the cultural stereotype of elderly people (e.g., wrinkle, bingo, Florida), these participants, despite being unaware of the elderly theme, walked down the hallway significantly more slowly than control groups. When the sentences contained a hidden message of rudeness (rather than politeness), the participants interrupted a social exchange more readily (Bargh, Chen & Burrows, 1996, Study 1). When the sentences contained a hidden message of cooperation (rather than competition), participants became more likely to cooperate than defect in a shared resource task (Bargh, Gollwitzer, Lee-Chai, Barndollar, & Troschel 2001). In fact, participant behaviors were influenced by information that is subliminal. Partici-

pants behaved significantly more angrily at a botched computer program if they had received subliminal exposures to pictures of African American. (Bargh, Chen & Burrows, 1996, Study 3). The main thrust of this research was that, even though people sense that they are fully in control over cognitive decisions such as whether interrupt a conversation, cooperate with a partner, express internal anger, or to rush off to an engagement, unidentified environmental forces can exert control over these behaviors.

Implicit Mental Representations

Further research demonstrated that people's decisions and behaviors can be governed by unidentified subconscious forces not just from the environment, but from their own minds. Measures of cognitive associations (such as the Implicit Associations Test ; Greenwald & Banaji 1995) measure people's spontaneous, unwanted, or subconscious associations to many concepts, and these associations do not necessarily parallel people's deliberative, espoused, or conscious beliefs. A classic line of research has investigated "aversive racism" – a situation in which a person sincerely subscribes to an egalitarian value system, yet harbors unwanted or unacknowledged negative feelings and beliefs towards blacks or other marginalized ethnicities (Dovidio & Gaertner 1986; Dovidio & Gaertner 2004). For example, many participants, regardless of their stated values, display significant quickness in recognizing words like "violent", "dangerous", and "lazy" after being briefly exposed to some representation of black people ((Fazio, Dunton, Jackson,

& Williams 1995) Supporting the notion that these associations are spontaneous and need not be consciously experienced, the "accessibility" of concepts like criminal, lazy, and violent occur even when the representation of black people is subliminal (Wittenbrink, Judd, & Park 1997).

Conflict Between Implicit and Explicit

Thus, a person's implicit associations and explicit beliefs may conflict (Greenwald and Nosek, 2009; Nosek 2005; Nosek 2007). In fact, the degree of conflict (vs. congruence) between a person's implicit and explicit cognition can itself be a predictor of important behavioral outcomes. For example, the degree of correspondence between a person's implicit and explicit goals predicts positive future health outcomes (Schultheiss & Brunstein 1999), such as daily experiences of emotional well-being (Brunstein, Schultheiss, & Grassman 1998). On the other hand, the degree of conflict between a person's explicit and implicit levels of self-esteem predicts anti-social behaviors such as outgroup denigration and violence (Jordan, Spencer, Zanna, Hoshino-Browne, & Correll 2003). The conflict may be unavailable to introspection, as studies have used subliminal conditioning to simultaneously create positive explicit attitudes and negative implicit attitudes towards the same novel object (e.g. Rydell and McConnell 2006).

The Potency of The Implicit

People's spontaneous or implicit associations seem to have importance for everyday life. Although the psychometric tests are merely artificial computer reaction time measurements, they predict meaningful social behavior (Greenwald, Poehlman, Uhlmann & Banaji 2009). For example, implicit outgroup associations can predict a hiring manager's decision to hire a Jewish vs. Catholic job applicant (Fein & Spencer 1997), an obese vs. non-obese job candidate (O'Brien, Latner, Halbertadt, & Hunter 2008), or a black vs. white job candidate (McConnell & Leibold 2001; Dovidio, Kawakami, & Gaertner 2002). Much research has supported the predictive validity of implicit measurements of associations to all sorts of concepts – races, genders, sexual orientations, body types, self-images etc. For example, implicit association tests have predicted nurse's decisions to make major career changes (von Hippel, Brenner & von Hippel 2008), pilot's tendencies for risk-taking behavior while navigating aircraft (Molesworth & Chang 2009), suicidal people's likelihood of attempting suicide in the next six months (Nock & Banaji 2007), doctor's biases in prescribing heart treatments (Green, Carney, Palin, Ngo, Raymond, Iezzoni & Banaji 2007), and future political support for enlarging a U.S. military base in Italy (Galdi, Arcuri, & Gawronski 2008)

1.1.2 The distinction between implicit and explicit cognition

In general, the experimental findings from the automaticity revolution in social psychology have been taken to support a distinction between implicit and explicit cognition.

By implicit cognition, social psychologists loosely refer to mental processes or contents that the person does not consciously experience, desire, or control, but which nevertheless influence social perception, decision-making, or behavior. More specifically, John Bargh (1994) emphasized that, in idealized form, implicit cognition refers to mental processes that fail to involve (a) intentionality, (b) controlled alterations, (c) subjective awareness, and (d) cognitive resources. That is, implicit cognition refers to the set of mental processes which an individual (a) does not intentionally instigate or "start up"; (b) cannot eliminate, alter, or override once started; (c) does not subjectively experience insight into their origins, meanings, or occurrence; and/or (d) does not spend limited cognitive resources to execute (Bargh 1994; Nosek 2007). For example (Sloman 1996), our visual system receives information from the retina and assembles a visual image very quickly and efficiently, without requiring the four features described above.

By explicit cognition, social psychologists loosely refer to those familiar, effortful mental processes that are subjectively experienced and open to our voluntary influence. With respect to John Bargh's (1994) four dimensions, cognition is said to be explicit which an individual (a) influences the process's instigation or "start

up”; (b) may eliminate, alter, or override the process when motivated to do so; (c) experiences some subjective awareness into the process’ origins, meanings, or occurrence; and/or (d) experiences depletion of limited cognitive resources because of the occurrence of the process. For example, the process of reasoning through a calculus problem or an LSAT exam generally is considered to require the four features of explicit cognition.

1.2 Theoretical Perspectives on Automaticity

1.2.1 The dualistic perspective

Based on the automaticity results, social psychological theorizing frequently draws upon dual systems theory, which holds that the human mind has two distinct computational systems with qualitatively distinct operating principles (Strack & Deutsch 2004; Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006; Rydell, McConnell, Mackie,& Strain, 2006; Smith & Decoster, 2000): one system is implicit, associative, uncontrolled, fast, subconscious, and unintended (System I; Kahneman & Frederick 2002), and the second system is explicit, rule-based, controlled, slow, conscious, and unintended (System II).

Dual systems models emphasize the dissociations between the two systems. The empirical goal is characterizing the systems’ different inputs and outputs. For example, studies have revealed that implicit attitudes are influenced by en-

vironmental conditioning, and influence spontaneous behavior, whereas explicit attitudes are influenced by verbal reasoning, and influence deliberative behavior (Rydell & McConnell 2006; Strack & Deutsch 2004). The theoretical explanations of behavior is also rooted in dualistic mechanisms, distinguishing between separate implicit and explicit entities. For instance, when person deliberately reports an explicit attitude rather than an explicit attitude, the mind is believed to retrieve a distinct representation stored separately in explicit memory (Wilson, Lindsey & Schooler 2000), or override simple associative network processing with symbolic rule-based computation (Strack & Deutsch 2004), or trigger a cognitively demanding response selection process which selects the personally endorsed mental representation from a set of automatically activated representations (Devine 1989).

However, because of this emphasis on dissociations, it remains unclear how the two systems would communicate with each other. In particular, because the explicit and implicit mind are believed to have separate computational forms – symbolic logical production rules vs. simplistic primitive associations (Sloman 1996; Fodor & Pylyshyn 1998; Pinker 1997) – it is not currently understood how these forms could interact (Greenwald & Nosek 2009). Yet there are numerous empirical results reflecting fluid communication between the implicit and explicit mind, in both directions. Research in evaluative readiness has demonstrated that a deliberately adopted strategic goal – such as the desire to win a game or to perform well academically – can moderate implicit attitudes within a matter of milliseconds (Ferguson & Bargh 2004; Ferguson 2008). Conversely, research on attitude conditioning has demonstrated that subliminal conditioning to novel ob-

jects influences people's explicit attitudes towards those same objects (Olson & Fazio 2001; Monahan, Murphy & Zajonc 2000).

1.2.2 The dynamical perspective

One primary goal of this dissertation is develop an understanding of the implicit and explicit mind from within a single dynamical framework, where interactions and communication is possible. To do this, we draw upon dynamical systems models of mental processing (Spivey 2007; Port & Van Gelder 1995; Kelso 1995), which emphasize that interactions underly the formation of mental representations. Dynamical models see explicit decisions as emerging out of the interactions of many components in a system. These models would differentiate between implicit and explicit cognition through a notion of "embedded knowledge" – that is, knowledge is embedded within many synaptic connections in a neural system, and mental processing requires neurobiological interactivity in order to generate an explicit representation. For example, when face recognition centers have strong training and sufficiently long processing time, neurobiological interactions can produce success on explicit measures; otherwise, the system succeeds only on standard measures of implicit cognition, such as priming and savings in re-learning (O'Reilly, Vecera & Farah, 1993). The corresponding notion that social judgments and decisions gradually emerge out of interactions of many components has been described by researchers applying dynamical perspectives to social psychology (Vallacher, Read & Nowak 2002; Vallacher & Nowak 1994; Freeman

& Ambady 2009; Freeman, Pauker, Apfelbaum & Ambady 2010; for similar approaches, see Judd, Drake, Downing, & Krosnick, 1991; Adolphs 1999; Fogel et al. 1992; Lewis, Douglas, Mascolo & Griffin 1998). These researchers have found evidence, for example, that group norms and individual self-esteem are the results of dynamical processes extended over time, in which the mind attempts to satisfy multiple constraints embedded within the initial states of the system (Vallacher, Nowak, Froehlich & Rockloff 2002). Moreover, neuro-imaging research on attitudes has demonstrated that mental representations gradually evolve from early amygdala-driven evaluations into increasingly more nuanced reflective attitudes over the course of recurrent processing dynamics in prefrontal-subcortical circuits (Cunningham & Zelazo 2007; Cunningham, Zelazo, Packer & Van Bavel 2008).

Because dynamical interactions may underly the formation of an explicit judgment or decision, the active constructivist approach to social decision-making becomes particularly useful (Schwartz 1994; Ferguson & Bargh 2007). The active constructivist perspective holds that a person's mental representations are not static, but are constantly reassembled anew upon recall, depending upon ever-changing contextual cues. For example, people's self-reported life satisfaction, which might seem to be a consistent stable mental representation, can vary as a function of amount of sunshine outdoors (Schwarz 2007). The active constructivist perspective therefore rejects the traditional file-drawer view of accessing mental representations from memory (see Klein, Sherman, & Loftus 1996); while it might seem to be the case that a person has memorized answers to certain ritualized questions (e.g. "Are you satisfied with your life?"; "Did you see the car crash?"),

instead the mind seems to be actively constructing mental representations on the spot. The context-dependence of social mental representations (attitudes, self-esteem, etc.) is quite consistent with the notion of dynamical interactions driving mental representations. The dynamical approach would suggest that these constantly reassembled mental representations are explained by online dynamical interactive processing that, at the moment of each social report, is attempting to best satisfy many simultaneously existing constraints. Thus, one moment's statement that "I like Republicans" may involve quite different dynamics than another moment's statement that "I like Republicans," rather than each report calling up a statically stored representation.

These considerations suggest that social decision-making may not be so different from perceptual decision-making, which is often described through dynamic conflict resolution processes. For example, the gradual formation of a phonemic representation over the course of hundreds of milliseconds (Did I just hear you say "bah" or "pah"?) can be modeled as the dynamic competition between multiple potential phonemic representations receiving varying degrees of support from voice onset time, aspiration, and over fourteen other acoustic features (McMurray, Tanenhaus, Aslin & Spivey 2003). In a similar manner, it may be the case that an explicit social judgment dynamically assembles from a large set of constraints or biases. Importantly, some of these constraints could be implicit biases, where implicit biases refers to "unwanted" mental contents (social beliefs, views, and associations) that could have been formed by any cognitive mechanism producing unintentional acquisition. Such mechanisms of unintentional acquisition

might include; unrecognized subliminal conditioning (Olson & Fazio 2001, Olson & Fazio 2002, Karpinski & Hilton 2001); persistent influences from rejected propositions (Petty, Brinol & DeMaree 2007; Smith & Decoster 2000); cognitive biases in the mind's inferential system, such as illusory correlations between the rare members of two categories (Hamilton & Gifford 1976; Hamilton & Rose 1980); or motivated attempts by the "psychological immune system" (Gilbert, Pinel, Wilson, Blumberg & Wheatley 1998) to restore damaged self-esteem by altering cognitive beliefs (see Spencer & Fein 1998; Sinclair & Kunda 1999; Balcetis & Dunning 2008). Through such mechanisms, a person may harbor mental contents matching the characteristics of implicit cognition; that is, a person may store mental contents that the person may not have intentionally acquired, may not necessarily endorse, may not be capable of controlling, and may not even be aware of – yet these mental contents could still potentially influence the dynamic formation of an explicit social decision or judgment.

Thus, in this dissertation, we seek to investigate the real-time dynamics of forming an explicit social judgment or decision, where real-time dynamics refers to the moment-to-moment timescale along which a person gradually forms that judgment or decision. In particular, we investigate the potential influence of implicit information sources, meaning mental contents formed by mental processes (such as subliminal conditioning, lingering rejected propositions, illusory correlations, or motivated cognition) which are devoid of intentionality, controlled guidance, subjective awareness, and/or cognitive resources. As argued above, there is some basis for believing that implicit information sources may be influencing the

real-time dynamics of forming an explicit social judgment. To summarize, the active constructivist view holds that explicit social decisions are actively constructed on the spot, the dynamical systems view holds that dynamical interactions underly those explicit social decisions, and the automaticity perspective holds that subliminal or unrecognized sources influence explicit judgments. Yet previous research has focused on the influence of how *static* implicit manipulations influence *static* explicit measures, and it remains unclear whether (and how) subliminal or otherwise implicit informational sources may be influencing unfolding dynamical PROCESS of explicit social cognition.

1.3 Empirical Goal: How do implicit biases influence the dynamics of forming an explicit judgment?

In chapters 2 and 3 of this dissertation, we depart from dual systems theory and its barriers between the implicit and explicit mind; instead we consider the formation of an explicit attitude as the end result of a dynamical process, extended over time, within a single interactive system. The term single interactive system is meant to encompass an influential class of neural network models such as Normalized Recurrence (NR; Spivey 2007), Simple Recurrent Network (SRN; Elman 1990), Dynamic Field Theory (DFT; Erlhagen & Schoner 2002), and Leabra (Leabra; O'Reilly & Munkata 2000). These models attempt to explain mental functioning with a single system deploying a single set of operating principles. This

endeavor would seem to be at odds with current dual system models in social psychology. This is because, if we examine the influential single systems models that have been developed in the broader cognitive sciences over the past 20 years, we see that these models would certainly get classified as "associative networks" rather than "rule-based systems." The single interactive system models are all constructed as parallel subsymbolic networks, rather than through serial symbolic logical rules. They all accommodate probabilistic soft constraints, rather than logical hard constraints. Thus, the single interactive systems models would match the description of an associative network.

However, while associative systems are assumed by dual systems theory to be primitive and unsophisticated, single interactive system models should be thought of as an "associative system" on steroids. The cognitive capacities of these single interactive system models far exceed those of traditional associative systems. In most dual systems models, the associative systems resemble various spreading activation networks developed in the 1970's (e.g., Collins & Loftus, 1974). However, the "connectionist revolution" in the 1980's and the "dynamic revolution" in the 1990's have advanced knowledge of what a "merely" associative system – that is of what a parallel distributed processing network – can do. Thus, here we coin the term "single interactive system models of cognition" to refer to these kinds of models (such as DFT, NR, and SRN, and LEABRA), with the new term reflecting the fact that these parallel distributed processing networks are no longer the associative networks of the 1970's, but self-organizing dynamical networks with sophisticated capabilities.

In particular, the connectionist revolution in the 1980's took the Collins and Loftus (1974) notion of associations between discrete symbolic concepts ("dog," "cat", etc.), and extended it to subsymbolic distributed representations (patterns of activation which may partially resemble several different concepts at once). The dynamic revolution of the 1990's took the move one step further by adding cyclic recurrent processing dynamics. Now, the layers of a connectionist network could interactively communicate with each other repeatedly over time. After these developments, it was possible to think about cognitive decisions, motor behaviors, or internal representations not as static, but as dynamically evolving over time. With the new features brought on by the connectionist and dynamic revolutions, single interactive system models began to exhibit quite sophisticated behavior. Researchers have shown that feedforward connectionist networks (FN) can perform deductive syllogisms (Rogers & McClelland 2004), that simple recurrent networks (SRN) can produce rule-based language (Elman 1990; Christiansen & Chater 1999; Bechtel and Abrahamsen 1991); that normalized recurrent networks (NR), can do serial-like visual search (Spivey 2007); and that dynamic field theory (DFT) can explain the development of "symbolic thought" (Thelen & Smith 1994; Schutte & Spencer 2002). Thus, these single interactive system models possess cognitive capacities that far exceed those of the original associative networks of the 1970s, even though these models are still parallel distributed processing networks, and therefore use operating principles that conform precisely to what dual systems models in social psychology would call "System I" rather than "System II."

How much can these single interactive system models explain? In Chapters 2 and 3 of this dissertation, we demonstrate that a single interactive system framework can be used to explain an apparently System II phenomenon – explicit attitudes (see Wojnowicz, Ferguson, Dale, & Spivey, 2009). In particular, we rely upon a single interactive system framework to describe how an explicit attitude gradually assembles from simultaneous interaction and rivalry among multiple implicit and explicit biases. From this perspective (in particular, following the NR model), the active construction of an explicit attitude occurs in the following way. In the preliminary moments of processing a stimulus during an evaluation task (e.g., “do you like or dislike Black people?”; “do you like or dislike Joe?”), a set of informational sources simultaneously provides graded probabilistic support for multiple potential explicit decisions (see Table 1.1 - still a). However, these informational sources continuously cascade probabilistic information to an integrative decision-making region (see Table 1.1 - still b). The integrative decision-making region accumulates evidence for candidate decisions, forces the potential decisions to compete by way of mutual inhibition (Chelazzi & Miller 1993), and then sends top-down recurrent feedback to the informational sources, thereby updating each source’s level of probabilistic support (see Table 1.1 - still c). This cyclic process reiterates many times, and over recurrent cycles of activation propagation, the system gradually resolves multiple simultaneously conflicting biases, thereby settling into a finalized conclusive stable representation (see Table 1.1 - still d). Behavioral evidence from chapters 2 and 3 suggest that this model does a good job of describing the real-time construction of an explicit attitude (for corroborating

evidence see Freeman, Ambady, Rule, & Johnson 2008), using parallel distributed processing rather than discrete symbolic logical rules.

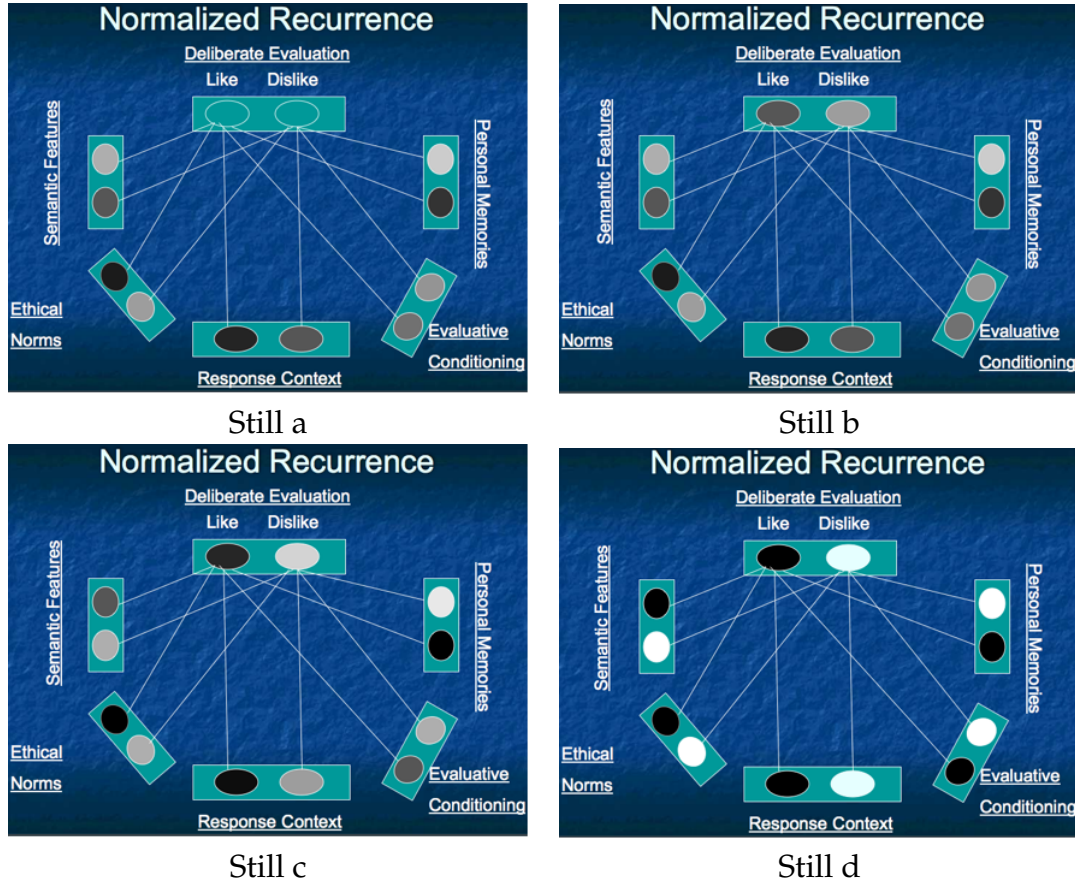


Table 1.1: A normalized recurrence model of a dynamically self-organizing deliberative evaluation. A node's activation is represented by darkness.

1.3.1 Chapter 2: The influence of aversive racism

In Chapter 2, we investigate a classic case of implicit-explicit conflict – racial attitudes. Studies on attitudes tend to find low correspondences between implicit and explicit measures of racial attitudes (typical correlations are between 0 and .33; Cunningham, Preacher & Banaji 2001; Dovidio, Kawakami & Gaertner 2002; McConnell & Liebold 2001), whereas implicit and explicit measures can correspond much more highly in other domains (political views such as pro-choice vs. pro-life attitudes can have correlations of about .70; Nosek 2005). In general, participants seem to have roughly equivalently positive explicit attitudes towards both white people and black people, but they seem to have much more negative implicit attitudes towards black people than to white people. From the dynamical perspective, these findings suggest that the seemingly identical explicit liking judgments towards black people and white people are merely the end results of very different dynamical processes in which there is much greater dynamical conflict during the real-time dynamics of formulating the liking judgment of black people. In particular, when participants are forming a liking judgment of black people, implicit informational sources may be providing particularly strong partial support for an unchosen dislike explicit attitude. Chapter 2 studies hand-movement trajectories as a means of tapping the real-time dynamics of forming a liking judgment. We find evidence that the process of explicitly evaluating black people involves elevated levels of dynamical conflict, with the conflict manifesting itself within three features of the dynamics (spatial deviations, strained velocity profiles, and

disorderliness).

1.3.2 Chapter 3: The influence of subliminal conditioning

In Chapter 3, we attempt to show directly that implicit (rather than explicit) informational sources can influence the dynamical process of formulating an explicit social decision. While the studies in Chapter 2 provide evidence that the real-time process of evaluating black people compared to white people involves greater dynamical conflict, there is no guarantee that such conflict was produced by implicit informational sources. People may simply possess more conflicting beliefs, views, feelings, and memories in the explicit realm, and the conflict distributed across those explicit informational sources may driving the elevated conflict expressed during the explicit report. To more directly test whether the elevated conflict traces back to implicit, subconscious informational sources, we employed a subliminal conditioning paradigm in the studies of Chapter 3. Using subliminal conditioning, we trained participants to have positive explicit attitudes towards two novel characters, but to have negative implicit attitudes towards one character and positive implicit attitudes towards another character. We then tested whether the subliminal conditioning exerts an influence on the dynamics of forming a liking judgment, even though participants remain completely unaware of the conditioning. The results provide some supporting evidence that this is the case.

1.4 Theoretical Goal: How does a dynamic mental system accomplish psychological control?

Psychological control can be defined as the ability to select thoughts and actions in the service of internal goals (Koechlin & Summerfield 2007). Since the previous empirical chapters have been concerned with making explicit responses that grate against implicit biases, it could be said that these explicit responses must have required some "psychological control." In fact, a good deal of research (reviewed in the introduction to Chapter 4) suggests that overcoming unwanted implicit biases requires cognitive control. For example, behavior is driven more strongly by implicit attitudes when cognitive resources are low. Moreover, the prefrontal cortex, typically described as the center of cognitive control, becomes activated at higher than baseline levels when people must override prepotent or habitual tendencies. But these observations raise a whole host of theoretical questions. In social psychological theory, the predominant models of psychological control are dualistic: there is a mental system which does the controlling (and that mental system, depending on the particular theory, might be characterized as rule-based, explicit, and/or rational), and a system which must be controlled (and that mental system, depending on the particular theory, might be described as associative, implicit, and/or emotional). Yet the controlled system is generally "outside" of the interactive dynamics, in the sense that dynamical properties tend to be stipulated and described only for the other (associative, implicit, intuitive, etc.) system. So then what, from the perspective of an interactive system, is psychological control?

Can we develop a full theory of control within an interactive system, a theory where we can understand how the primary capacities of a controller could arise out of dynamical interactions in a parallel distributed system like the brain?

1.4.1 Chapter 4: A dynamical interactive model of psychological control

In Chapter 4, we sketch a theory of psychological control within a dynamical interactive system. First, we explore dualistic accounts of control to uncover three primary capacities accorded to the privileged control system – decisive selection, the strategic pursuit of a distant goal, and flexible toggling between goals. In dualistic theories, these capacities are seen to be impossible by the more primitive system, which tends to be the system which is distributed and dynamic; thus, such theories make it unclear how a dynamic, distributed system would accomplish such feats. However, we briefly argue, as a point of departure, that any reasonable theory of control must be built out of interactivity. Thus, we begin by assuming that any theory about the brain must adhere to four common principles of biologically plausible mental processing, and based on these principles, we attempt to discover how the three control capacities would work from within such an interactive system. We will discover that in the end, we indeed need multiple specialized systems to perform psychological control, but that computations within each of these specialized systems take the form of subsymbolic parallel dis-

tributed processing. For the reason, the mechanisms for interactions both within and between systems can be made conceptually clear, both at a biological level (i.e. synaptic connections) and a mathematical level (i.e. coupled firing rates). The primary result of this chapter is to describe an "interactive" biologically plausible realization of the three control capacities. We conclude the chapter by proposing a crisp geometric hypothesis about how psychological control works in an interactive brain, and we apply the hypothesis to an important case of psychological control of the implicit mind, namely evaluative readiness.

CHAPTER 2

THE DYNAMICAL INFLUENCE OF AVERSIVE RACISM ON EXPLICIT EVALUATIVE PROCESSING

2.1 Overview

Note: This chapter has been published under the alternate title: The self-organization of explicit attitudes.

How do minds produce explicit attitudes over several hundred milliseconds? Speeded evaluative measures have revealed "implicit" biases beyond cognitive control or subjective awareness, yet mental processing may culminate in an "explicit" attitude that feels personally endorsed and corroborates voluntary intentions. We argue that self-reported "explicit" attitudes derive from a continuous temporally dynamic process, whereby multiple simultaneously conflicting constraints self-organize into a meaningful mental representation. As our participants reported their explicit (like vs. dislike) attitudes towards white versus black people, we recorded streaming x, y coordinates from their hand-movement trajectories. We found that when participants reported positive explicit attitudes toward black people, rather than white people, their hand movement paths exhibited greater curvature toward the dislike response. Moreover, these trajectories were characterized by precisely the movement disorder and competitive velocity profiles predicted under the assumption that the deliberate attitudes were

emerging from continuous interactions between multiple simultaneously conflicting constraints.

People's evaluations of stimuli as positive or negative can be activated in memory in either an unintentional (implicit) or intentional (explicit) fashion (e.g., for reviews, see Albarracin, Johnson, & Zanna, 2005; Fazio & Olson, 2003; Ferguson, 2007; Petty, Fazio, & Brinol, 2007; Wittenbrink & Schwarz, 2007). That is, people's attitudes toward stimuli can at times be rapidly and unintentionally activated by the mere presence of the stimuli, as well as deliberately and intentionally reported. Research in social cognition over the last two decades has shown that people's implicit and explicit attitudes toward the same stimulus are sometimes dissociated, particularly when the stimulus represents a socially stigmatized group or person (e.g., Nosek, 2005). For example, although very few participants explicitly report more positive attitudes toward white versus black people, such a pro-white preference nevertheless emerges in many participants' implicit attitudes. Participants shown computer-based images of black versus white people are significantly slower in making lexical decisions about subsequently encountered positive words (e.g., sunshine, puppy) and faster at making decisions about negative words (e.g., disaster, cancer), which suggests they evaluated the images of the black people in a relatively more negative manner (e.g., Fazio, Jackson, Dunton & Williams 1995; Wittenbrink, Judd & Park 1997). Given that many people show initial, implicit biases toward traditionally stigmatized groups in society, how do they overcome them when explicitly reporting positive attitudes toward the same groups? In other words, how do people generate intentional attitudes, especially

those that involve potentially conflicting sources of influence?

A variety of theoretical accounts has been proposed to accommodate existing data on the formation of attitudes and choices, ranging from the broadly framed dual-attitude models (Wilson, Lindsey, & Schooler, 2000) to the more specific dual-process models (Devine, 1989; Smith & DeCoster, 2000) to dynamic interaction models (Judd, Drake, Downing, & Krosnick, 1991; see also Roe, Busemeyer, & Townsend, 2001, and Heider, 1946). What we find in common in all of these accounts is the co-existence of multiple attitudes and an emphasis on the temporal dynamics of how they influence evaluative responses. Rather than selecting among the specific theories, we invoke the encompassing theoretical framework of self-organization to guide an exploration of those temporal dynamics, with specific predictions for what should result from multiple attitudes interacting over time.

Starting from the premise that mental representations in general are dynamically evolving states (Conrey & Smith, 2008), we suggest that explicitly reportable attitudes are merely the end result of a complex nonlinear time-dependent process of multiple less-explicit attitudes competing with one another over hundreds of milliseconds. Mental representations, as implemented in the brain, are distributed representations: neural populations convey information through patterns of firing rates distributed across multiple neurons (Rogers & McClelland 2004; Spivey, 2007), even in higher-order decision-making regions (Bogacz & Gurney, 2006; Lapish, Durstewitz, Chandler & Seamans, 2008). Contemporary decision-making

frameworks therefore model the decision-making process as a dynamic real-time evolution of a distributed pattern (Busemeyer & Townsend, 1993, Usher & McClelland, 2003). In early moments of mental processing, these distributed patterns are partially consistent with multiple interpretations, due to their proximity to multiple neural population codes, and they therefore provide a vague, preliminary interpretation. However, a continuous accrual of information causes the distributed pattern to dynamically sharpen into a confident (selected) interpretation, forcing other partially activated competing alternative representations, decisions, or actions to gradually die out. Thus, in this self-organization framework, it is possible for one attitude (whose biases rise quickly in activation) to be briefly prominent during early moments of forming an attitude choice, whereas during later moments of forming that same attitude choice, a different attitude may take hold (whose biases are stronger but rise in activation more slowly). That latter attitude will eventually activate other subsystems, such as language and memory, thus making the attitude seem explicit. What makes the first attitude implicit is not necessarily that it was generated in a different subsystem, but simply that it did not hold sway long enough to activate those language and memory subsystems.

This basic framework places cognitive processes in the same domain as many other natural phenomena that evolve through self-organizing dynamics (Kelso, 1995; Van Orden, Holden & Turvey, 2003). Self-organizing systems change states over time, even under constant input: these systems are endowed with internal constraints which cause them to change states autonomously, because continu-

ous interactions between component parts (e.g. population codes for an interpretation or behavioral choice) drive that system through a series of intermediate states towards a stable steady state. In the mind, processing generically involves recurrent processing loops (or cyclic feedback) between higher-order integrative regions and lower-level informational sources (Lamme & Roelfsema, 2000; O'Reilly, 1998, Spivey, 2007). Moreover, these higher-order integrative regions enforce representational competition, whereby increasing the activation of one particular interpretation inhibits alternatives (e.g. Miller 2000; Desimone & Duncan 1995). In this way, highly probabilistic mental states morph into nearly discrete symbolic representations. Many behavioral studies have supported that higher-order mental states continuously evolve through the dynamic satisfaction of multiple simultaneously conflicting constraints, even for seemingly categorical decisions in speech perception (McMurray, Tanenhaus, Aslin & Spivey, 2003), syntactic rule construction (Farmer, Anderson, & Spivey, 2007), and semantic categorization (Dale, Kehoe & Spivey, 2007). In the present work, we extend this framework to self-reported attitudes regarding social preferences.

A self-organizing, explicitly reported attitude requires a set of informational sources, including, for example, semantic features, evaluative conditioning, personal memories, motivational value, and response context. These informational sources should continuously cascade intermediate results of processing into integrative decision-making regions, such as the basal ganglia (Bogacz & Gurney, 2007) and cortical motor areas (Cisek & Kalaska, 2005). These sources send simultaneous probabilistic support for multiple candidate decisions; so in early

moments of processing, semantic knowledge might be 70% supportive of a like decision and 30% of a dislike decision. Higher-order integrative regions force the potential evaluative representations to compete, and then send top-down recurrent feedback to the informational sources. Gradually, through multiple cycles of recurrent processing, the system self-organizes into a coherent response (Spivey, 2007). From this perspective, the research showing pro-white biases in peoples implicit attitudes (e.g., Fazio et al., 1995) suggests that the black people stimulus may evoke greater conflict distributed across probabilistic information sources as the positive deliberate evaluation dynamically emerges. If so, a temporally fine-grained analysis should reveal that peoples explicit liking judgments for black people and white people evolve in real-time processing with qualitatively different dynamics.

How might we capture such dynamical information in real time? The unfolding cognitive dynamics may be revealed in continuous motor output. Because mental processing is recurrent, motor representations begin specifying movement parameters probabilistically rather than waiting for a perfectly completed cognitive command (Erlhagen & Schoner, 2002). In fact, motor commands may initiate movement before specifying a unique target destination, because motor trajectory parameters can be continually updated mid-flight (Henis & Flash 1995). For example, manual reaching towards a verbally named target object (e.g. candy) curves more towards a distractor that has a similar sounding name (e.g., candle; Spivey, Grosjean & Knoblich 2005). Manual reaching towards an animals taxonomic classification (e.g. mammal) curves more towards the distractor response

(e.g. fish) for taxonomically equivocal animals (e.g. whales vs. apes; Dale, Kehoe & Spivey 2007).

The motor execution of explicitly reported attitudes toward different ethnic groups may exhibit similar nonlinear dynamics. To test whether explicit attitudes toward potentially conflicting stimuli show such competition, we tracked participants motor trajectories towards like/dislike responses representing their explicit attitudes. Given the implicit attitude findings concerning black versus white people, and assuming that explicit attitudes dynamically emerge through self-organization, we predicted that hand trajectories should show greater motor curvature towards a dislike response while participants are positively evaluating black rather than white people. This motor curvature would reveal a greater influence of a dislike decision during the process of settling into a like decision toward black people. Additionally, there are two fine-grain predictions that are made exclusively by a competitive dynamics account of this phenomenon. Assuming that the black people stimulus evokes elevated dynamic competition between simultaneously active like and dislike representations, movement trajectories for the black people stimulus should exhibit evidence of nonlinear dynamics in their velocity profiles, as well as increased spatial disorder in the curviness of the trajectories.

2.2 Study 1

2.2.1 Study 1: Methods

Streaming x and y coordinates of mouse-cursor movements were recorded from 68 Cornell University undergraduates (43 female) in a simple explicit attitude task. Trials began with two seconds for participants to view the evaluative response options (LIKE) and (DISLIKE) randomly assigned to upper corners of the screen. Participants then clicked on a small box at the bottom of the screen to reveal a stimulus word and dragged the mouse toward their selected evaluative response. The 40 stimulus words included the target stimuli, black people and white people, as well as 19 positively valenced distractors (e.g. sunshine, babies) and 19 negatively valenced distractors (e.g. rats, murderers). These 40 stimulus words were presented in two blocks with randomly assigned order. The two stimulus repetitions were averaged together to yield a single measurement per participant for all statistical analyses (but not distributional analyses). We analyzed data only from the 61 participants who reported liking both white people and black people on both stimulus repetitions.

2.2.2 Study 1: Results and Discussion

Compared to the white people trajectories, the black people trajectories curved significantly more toward the dislike response option, as shown in Figure 2.1 (upper half). The maximum perpendicular deviation from a hypothetical straight line connecting the trajectories start and endpoint was greater for the black people trajectories than for the white people trajectories, $t(60)=2.17$, $p\text{-rep}=.94$, $d=.29$. As a result, the mean distance traveled en route to the like response was also longer for black people trajectories than for white people trajectories, $t(60) = 2.44$, $p\text{-rep}=.98$, $d=.32$. Responses did not differ in total reaction time, $t(60)=1.44$, $d=.18$.

In principle, the observed differential motor curvatures could be generated by a stage-based sequence of decisional commands, rather than by continuous motor attraction to the dislike response. If motor execution required the complete pre-specification of a unique target destination, rather than tracking motor trajectory parameters that continuously evolve mid-flight, then a mean trajectory could look differentially curved by averaging in replanned trajectories (where some proportion of trials involve an initial motor command guiding movement directly towards dislike that becomes aborted and replaced by a second motor command towards like). To accommodate the empirical mean trajectory that initially moves upward, rather than actually toward dislike, such an account would need to predict a bimodal distribution of curvatures, with some trajectories very curved and others not curved. However, the distribution of trajectory curvatures shows no evidence of bimodality, as shown in Figure 2.2. The degree of bimodality can

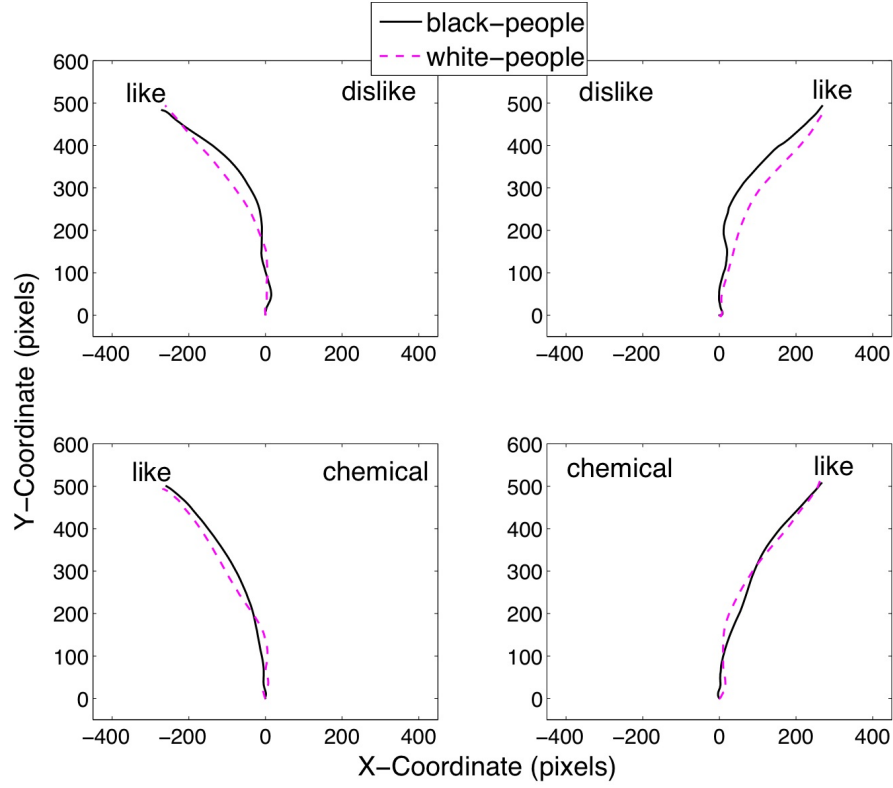


Figure 2.1: *Studies 1 and 2*: Mean mouse-movement trajectories towards evaluative targets for the black people and white people stimuli

be quantified with a bimodality coefficient, capable of detecting bimodality in a mouse-tracking paradigm (Spivey et al., 2005). The bimodality statistic is computed through the following formula (DeCarlo, 1997):

$$b = \frac{(skewness^2) + 1}{kurtosis + \frac{3(n-1)^2}{(n-2)(n-3)}},$$

where n is the number of observations in the distribution of interest. The standard cutoff for inferring bimodality in a distribution is $b_{.55}$. Neither the black- nor

white-people distributions of trajectories met this cut-off, and in fact the black-people trajectories form a distribution of movement curvature that is actually closer to normal, $b=.24$, skewness=.613, kurtosis=2.57, than the white-people trajectories, $b=.301$, skewness=.98, kurtosis=3.44.

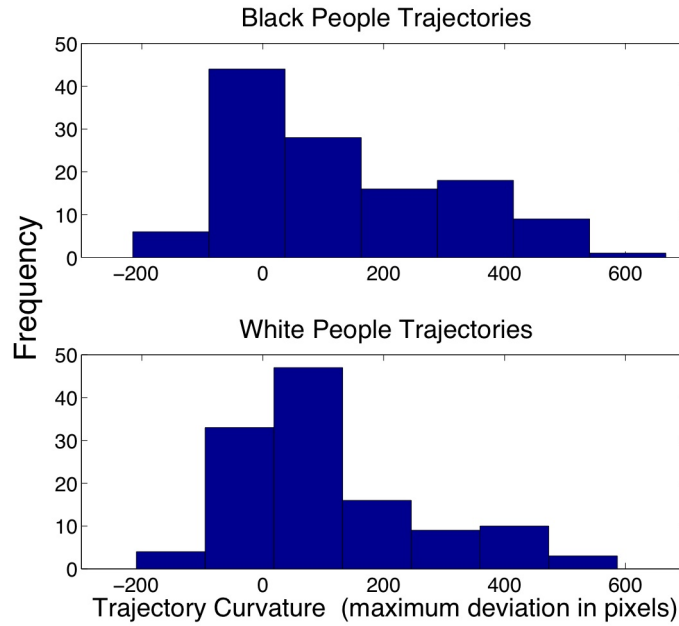


Figure 2.2: *Study 1*: Distributions of trajectory curvature show no evidence of bimodality

We further analyzed these computer-mouse trajectories for evidence of non-linear competitive dynamics, a signature of complex self-organizing systems. Velocity profiles were constructed by analyzing the temporal derivatives of motion towards the like response box along the x-coordinate. Movement along the x-coordinate axis reflects relative confidence in deciding upon one evaluation over

the other, since the mouse-movements starting location is equidistant between the two response boxes along this horizontal dimension. Our velocity predictions come from Usher and McClelland's (2003) differential equations for modeling the dynamics of competition between mental representations:

$$dx_1 = (I_1 - x_1 - \beta f_2)dt$$

$$dx_2 = (I_2 - x_2 - \beta f_1)dt$$

where in this case x_1 and x_2 represent the activations of the mental representations for like and dislike; dx_1 and dx_2 represent the change in the activation of the two mental representations in a time step of size dt ; I_1 and I_2 represent excitatory input to the representations from informational sources; βf_1 and βf_2 represent the inhibitory input of each mental representation to the other (called lateral inhibition); and f_i ($i=1$ or 2) is equal to x_i if x_i is greater than 0. According to these differential equations for competition dynamics, a strong evaluative competitor (x_2) sends intensified and prolonged lateral inhibition (βf_2) to the like evaluation (x_1). Thus strong competition alters the velocity profile of the movement toward the evaluative attractor ($\frac{dx_1}{dt}$), reducing velocity toward the attractor early on in processing. However, as the more active alternative begins to win the competition, this lateral inhibition is gradually lifted, thus increasing velocity later in processing to produce greater acceleration. Therefore, strong competition predicts higher acceleration into the like response box ($\frac{d^2x_1}{dt^2}$) in normalized time. Moreover, this particular dynamic pattern (reduced early velocity and greater later acceleration)

should lead to greater peak velocity, if jerk is minimized as the system achieves equivalent integral under the curve (representing net change in activation or location). Thus, dynamic conflict does not simply delay processing, but changes its composition: strong competition should lead to compressed high-spiking derivative profiles toward the preferred interpretation, even in normalized time.

These mouse trajectories approach the like response boxes with precisely the temporal derivative profiles predicted by Usher and McClelland's (2003) model of competition dynamics, as shown in Figure 2.3. The black-people trajectories had significantly greater maximum x-coordinate acceleration (shown as steeper velocity slope) into the like response box, than the white-people trajectories, $t(60)=2.69$, $p\text{-rep}=.96$, $d=.41$. Moreover, the black-people trajectories had significantly greater peak velocity (shown as higher velocity curve peak), $t(60)=2.65$, $p\text{-rep}=.95$, $d=.36$. These findings suggest that mental representations for both response alternatives, like and dislike, may be simultaneously active and competing over time, as in the Usher and McClelland model.

The spatial disorder analysis investigated the regularity of change in x-coordinate location over time. Our prediction about spatial disorder draws upon previous work on natural and physical self-organizing systems, which has established that increasingly conflicting constraints on a system's state invokes dynamic state-space trajectories that show more disorder or irregularity in their pathways (Kauffman, 1993; see also Dale et al., 2007, and McKinsty et al., 2008). In the present study, a self-organizing framework predicts that the motor trajec-

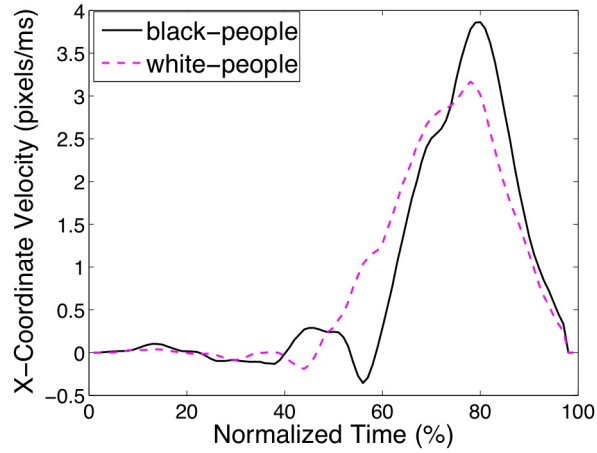


Figure 2.3: *Study 1*: Mean velocities towards the evaluative decision for the black people and white people conditions in normalized time

tories for the black people stimulus should have greater disorder than those from white people trials, even in the segment of the trajectory which has already committed to a like response. To investigate whether the black people trajectories have more wiggles, blips, and other such irregularities, we analyzed x-coordinate location over time, but only after the trajectory began moving in the positive x-direction. A sigmoidal fit, which very snugly fits curves that continue asymptotizing toward like in an orderly regular manner, as shown in Figure ??, was then imposed on the obtained curve. The black-people trajectories had significantly greater deviation from the sigmoidal fit, as revealed in a significantly lower R-squared, $t(60)=2.29$, $p\text{-rep}=.92$, $d=.31$, indicating disorderly variation around the x-dimension in those trajectories.

The curvature results (Figure 2.1) clearly demonstrate a greater motor attraction toward the dislike response option for black people trials, indicating some initial prominence of this negative evaluation in responses that, a fraction of a second later, manifest as positive attitude choices. It is worth noting that this difference in curvature emerges in the absence of a difference in total reaction time. The findings in the velocity and spatial disorder analyses further suggest that this initial prominence of the negative evaluation may be part of a dynamic process of parallel competition between partially active positive and negative implicit evaluations, the winner of which becomes the explicit attitude choice.

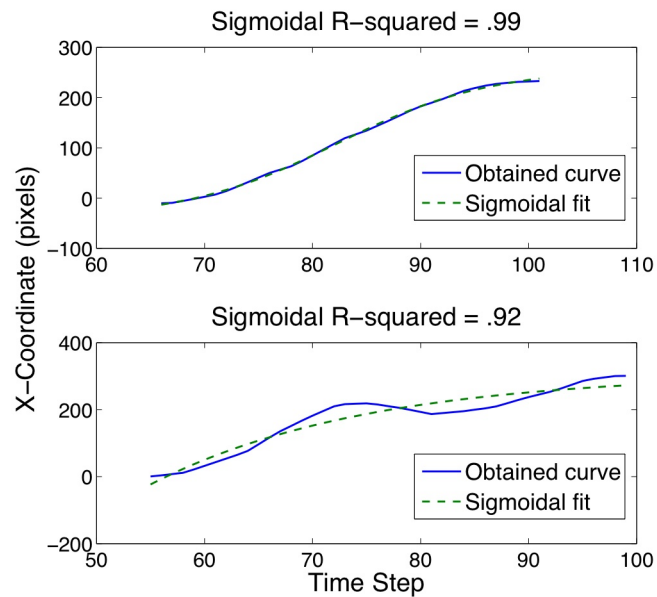


Figure 2.4: *Study 1*: Hand trajectories exhibit greater spatial disorder during evaluations of black people than white people

2.3 Study 2

As our claim is that multiple partially active mental representations compete for the privilege of driving evaluative responses, imposing a set of response options that are not particularly competitive should change the motor dynamics. If the response box opposite to the like box does not provide any semantic match to the content of the self-organizing evaluative response, then white people and black people trajectories should lose their differential curvature. In particular, the targets black people and white people should evoke much stronger support for interpretations as positive entities than as chemical elements.

2.3.1 Study 2: Methods

Sixty-six Cornell University undergraduates (40 female) were asked to classify words (e.g. ice cream, sunshine, boron) as something they liked ("LIKE") or as the name of a chemical element ("CHEMICAL"). We analyzed data only from the 63 participants who consistently reported "liking" both black people and white people on both repetitions of these trials, and who reported in a post-study questionnaire that they were not forced into selecting like by the paradigm.

2.3.2 Study 2: Results and Discussion

According to statistical analyses on maximum deviation and distance traveled, the black people and white people trajectories no longer differed in their curvature toward the competing response, as shown in Figure 2.1 (lower half), $t(62) = -.10$, $p\text{-rep}=.16$, $d = -.01$. Thus, the results of Experiment 1 are not merely due to responses to black people involving a longer latency to settle on a positive evaluation, and thereby drifting for longer in empty regions of movement space before curving toward the like response box. Rather, the dislike response option in Experiment 1 is actively pulling movement trajectories toward it, in a way that the chemical response option in Experiment 2 does not do.

There were no significant differences in maximum acceleration, $t(62)=-1.06$, $p\text{-rep}=.64$, $d=.17$ or peak velocity, $t(62)=-1.39$, $p\text{-rep}=.74$, $d=.22$, which if anything trended toward higher values for white-people trajectories. Likewise, there were no significant differences in spatial disorder, $t(62)= -.13$, $p\text{-rep}=.19$, $d = -.02$.

2.4 Study 3

Whereas we frame our results with respect to explicit attitudes toward people of different races or ethnicities, the mouse-cursor response trajectories to black people and white people in Experiment 1 may have diverged due to subtle confounds that do not refer to people at all. For example, perhaps these differences

reflect different evaluations of the color terms black and white that precede the term people.

2.4.1 Study 3: Methods

Seventy-one Cornell University undergraduates (37 female) were asked to classify stimuli as something they liked ("LIKE"), or disliked ("DISLIKE"). The methods were the same as Study 1, except that the crucial stimuli in this experiment were the terms African Americans and Caucasians, rather than "black people" and "white people." We analyzed data only from the 64 participants who consistently reported "liking" both African Americans and Caucasians on both stimulus repetitions.

2.4.2 Study 3: Results

The African-American trajectories curved significantly more towards the dislike response than Caucasian trajectories, $t(63)=3.65$, $p\text{-rep}=.99$, $d=.56$, as shown in Figure 2.5, which shows averaged rightward and horizontally-flipped leftward trajectories.

The motor trajectories temporally evolved in accordance with the competitive velocity predictions, as reported in Experiment 1. The African-American trajectories, compared with the Caucasian trajectories, had significantly greater max-

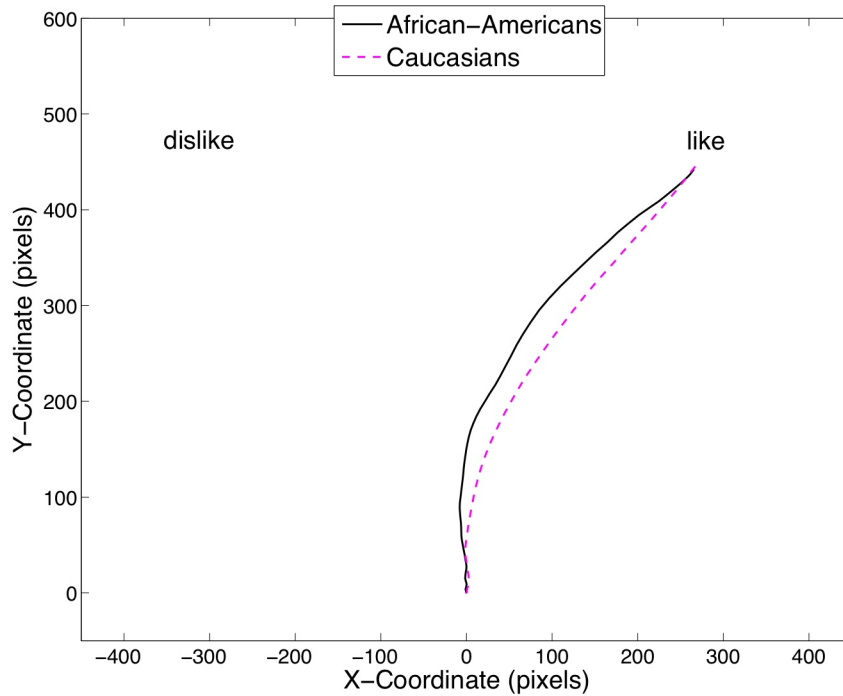


Figure 2.5: *Study 3*: Mean mouse-movement trajectories towards evaluative targets for the stimuli African Americans and Caucasians in the third study

imum x-coordinate acceleration, $t(62)=3.55$, $p\text{-rep}=.99$, $d=.47$. These trajectories also obtained higher peak velocity, $t(62)=4.54$, $p\text{-rep}=.99$, $d=.63$. Moreover, as seen for black people trajectories in Experiment 1, the African-American trajectories exhibited greater spatial disorder (than Caucasian trajectories), even once moving towards like, as indicated by significantly greater mean deviation from the sigmoidal fit, $t(62)=2.49$, $p\text{-rep}=.94$, $d=.44$. In tandem, these results demonstrate that the same general constellation of findings is observed with the labels African

American and Caucasian as was observed with the labels black people and white people.

2.5 Discussion

Peoples hand-movement trajectories for explicitly evaluating black versus white people distinguished themselves along three properties of the dynamics: shape, time, and order. These findings suggest that explicit attitudes evolve through continuous temporal dynamics during real-time mental processing, with graded motor curvature revealing the influence of dislike tendencies. There was no evidence for cleanly separated (i.e., discrete, rather than continuous) explicit decisions, whereby an initial response was executed solely toward the dislike response box and then a corrective response was executed mid-flight toward the like response box. Rather, the results suggest that a dynamic competition process may be what allows a single explicit attitude choice to emerge from multiple, potentially conflicting evaluative influences (e.g., Busemeyer & Townsend, 1993; Usher & McClelland, 2003). Thus, the mind, rather than switching from one singular (implicit) decision to a different singular (explicit) decision, may host a continuously evolving blend of (implicit) evaluative decisions from which the eventual (explicit) behavioral choice emerges.

CHAPTER 3

**THE DYNAMICAL INFLUENCE OF SUBLIMINAL CONDITIONING ON
EXPLICIT EVALUATIVE PROCESSING**

3.1 Overview

3.1.1 A "dual channels" model of evaluative decisions

The term "simultaneous contradictory belief" (Sloman 1996) has been coined to describe situations in which people simultaneously believe two contradictory responses. Examples include people's contradictory responses to the Muller-Lyer illusion, statements like "technically, a whale is a mammal," and separate pulls from similarity and logic in the Linda-the-bank-teller problem. In such situations, people report feeling their minds pulling them in two different directions at once.

Social psychology theory frequently suggests that, at a broader scale, people can be "Of Two Minds" (e.g. Smith & Decoster 2000, Gawronski & Bodenhausen 2006, Strack & Deutsch 2004). Seemingly, mental systems can clash – such as when rationality clashes with intuition, when a temptation gets the better of a person, or when cerebral and emotional selves disagree. In particular, the automaticity revolution has emphasized the potential clash between the implicit and explicit minds (e.g. see Chartrand & Bargh 1999; Berridge & Winkielman 2003). For example, dual attitude models (Wilson 2000; McConnell and Leibold 2001; Rydell &

McConnell 2006) hold that implicit attitudes and explicit attitudes are completely separable. They are trained by different kinds of information (e.g. subliminal vs. verbal) and they direct different kinds of behavior (e.g. controlled vs. uncontrolled).

Empirical research has been generated to support the existence of simultaneously contrasting beliefs at the implicit and explicit levels. For example, in one set of studies (Rydell & McConnell 2006) participants were trained to learn about a new person, Bob, over the course of 100 learning trials. For each trial, participants would see the picture of Bob, along with some verbal information potentially characterizing his behavior. These behavioral statements reflected behavior that was either clearly positive, as in "He donated a pint of blood to the Red Cross," or clearly negative, as in "He cheated on a take-home exam from the university." Participants were supposed to guess whether the sentence characterized Bob accurately or not. For participants in condition A, the good sentences always characterized Bob (and the bad sentences were always uncharacteristic). For participants in condition B, the bad sentences always characterized Bob (and the good sentences were always uncharacteristic). However, participants encountered directly contrasting information at the subliminal level. In condition A, the pictures of Bob were always preceded by a 25 millisecond negative subliminal prime (like "vomit," "death"), whereas in condition B, the pictures of Bob were always preceded by a 25 millisecond positive subliminal prime (like "flowers," "happiness").

The researchers found that such training leads to countervailing effects: par-

ticipants in condition A exhibited positive explicit attitudes but negative implicit attitudes, whereas participants in condition B exhibited negative explicit attitudes but positive implicit attitudes. Moreover, when the same participants were exposed to the opposite procedure (i.e., after testing, participants in condition A were exposed to the condition B training), the results reversed. The results are shown in Figure 3.1.

Based on these results, the researchers argue that people are “of two minds” and can simultaneously hold independent contradictory attitudes. These distinct attitudes were formed from distinct environmental inputs: explicit attitudes formed and changed in response to consciously accessible behavioral statements, whereas implicit attitudes formed and changed in response to subliminally presented primes. Moreover, based on other studies in the field (McConnell & Leibold 2001; Dunton et al. 2002), these distinct attitudes drive distinct behaviors: explicit attitudes drive the verbal contents of behavior, whereas implicit attitudes drive nonverbal communication. The idea is of two independent channels of mental processing.

The notion of two distinct channels leaves little room for notion of interactivity. The researchers conclude that their results “seem incompatible with models of evaluation that assume explicit attitudes are simply modified version of implicit attitudes accessed from memory.” However, the notion of strictly independent channels seems to grate against one of the mind’s major fundamental tasks – that information conveyed by specialized groups of neurons must be functionally in-

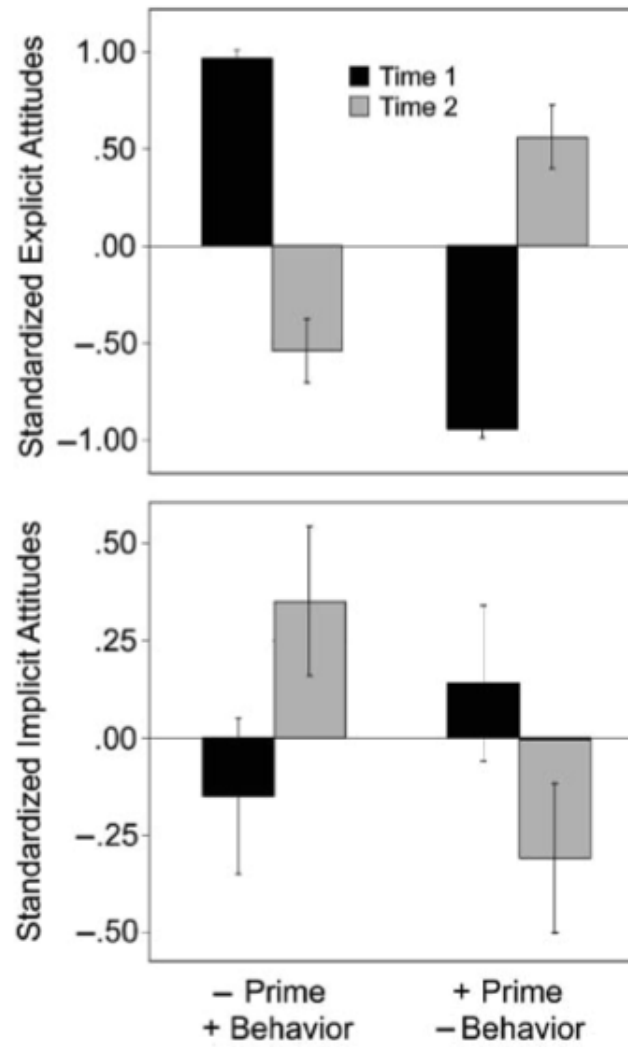


Fig. 1. Explicit (top panel) and implicit (bottom panel) attitudes as a function of condition and time. Error bars represent standard errors.

Figure 3.1: Primary findings of Rydell & McConnell 2006

tegrated in order to guide behavior (Tononi, Edelman, & Sporns, 1998). The mind would be rather impoverished if its learning from subliminal evaluative conditioning could not influence adaptive behavior – and in fact, evidence that this occurs has been obtained many times (e.g. Zajonc 1968; Nisbett and Wilson 1977). Thus, in this study, we ask whether it is possible that a more sensitive measure would reveal that explicit attitudes are actually tracking, at least in the early moments of real-time processing, the results of the subliminal conditioning?

The theoretical interest in this question comes from a single interactive systems approach to social decision-making (described in detail in Chapter 1). From this perspective, explicit mental representations, such as attitudes, gradually unfold in real-time processing as information sources dynamically self-organize into a coherent decision. This model would predict that if participants in a study were trained to like Bob, but participants in the cohort condition were exposed to subliminally negative associations and participants in the control condition were exposed to subliminally positive associations, then participants in the cohort condition would have more conflict distributed across information sources, and thus may exhibit greater evidence of dynamical conflict while making their decision.

This dynamical interactive approach to evaluative decision making was supported by the empirical results in the aversive racism chapter (Chapter 2). Because of the documented widespread existence of aversive racism, the chapter expected greater conflict during the real-time evaluative processing of black people than during the process of evaluating white people. The single interactive systems

model in turn predicts the degree of conflict initially distributed across information sources should manifest itself in the dynamic process of making an evaluative decision. The analysis of mid-processing hand-movement trajectories supported this view of greater dynamical conflict during the evaluations of black people. However the research in Chapter 2 failed to address exactly *which* informational sources are contributing to the evaluative dynamics. The dynamic formation of an explicit attitude towards different ethnicities could depend upon solely declarative information (semantic knowledge, personal memories, task context, etc), and not draw at all from subliminal conditioning. Perhaps, somehow, subliminal conditioning is simply too weak of an influence to alter explicit evaluation processes.

Thus, the Rydell and McConnell research raises a particularly interesting question: Can *subliminal conditioning* be an informational source to the dynamics of constructing an *explicit* evaluation? This is precisely the question we study in this chapter. Thus, in this study, we taught participants about two novel (fake) characters. One character was described through dissociated training (positive behavioral statements, negative subliminal conditioning), and the other character was described through consistent training (positive behavioral statements, positive subliminal conditioning). After the learning phase, we gave participants a two-alternative forced choice (LIKE vs DISLIKE) attitude task, and we tracked the hand-movement trajectories of the participants. We predicted that subjects would report liking both novel characters but that their hand-movement trajectories would curve more towards dislike for the dissociated stimulus than for the consistent stimulus. Such a result would show the real-time influence of sublim-

inal information on the early moments of explicit decisions. This finding would support the theoretical notion that explicit representations (here, attitudes) gradually unfold from the self-organization of implicit biases.

3.2 Pilot Study

3.2.1 Pilot Study: Methods

At a broad level, we trained participants using the following procedure

STIMULUS	ASSOCIATIVE PRIMING	STATED FACT
Joe	Bad	Good
Bob	Good	Good

Our experiments roughly followed the previously discussed procedure of Rydell and McConnell (2006). We trained participants through a "learning task" (actually an attitude induction procedure), whereby participants would learn about two novel characters, Bob and Joe. The pictures used for Bob and Joe were randomly selected from a bank of five pictures. For each character, subjects viewed 50 trials of behavioral statements that may potentially describe the character, such as "He donated a pint of blood to the Red Cross" or "He cheated on a take-home exam from the university." Participants were asked guess whether the statement is true or false. As it turns out, following Rydell and McConnell (2008), the positive statements are always true, and the negative statements are always false. Dur-

ing the learning task, we attempted to induce implicit attitudes towards the novel characters through a subliminal priming technique. The picture for one novel character (randomly selected, but we refer to him as "Joe") was always preceded by a negative subliminal prime (like "vomit," "death"). In contrast, the picture of the other novel character ("Bob") was preceded by a positive subliminal prime (like "flowers," "happiness").

More specifically, the learning task was implemented on a computer with a screen refresh rate of 85 Hz and a screen resolution of 1024x768 pixels. For each participant, the subliminally negative character was randomly assigned to be named either Bob or Joe, and the subliminally positive character was given the other name. Moreover, the pictures for each character were randomly selected from a bank of 5 pictures. The learning task overall comprised a bank of 50 pre-constructed trials for the subliminally positive character and 50 pre-constructed trials for the subliminally negative character. As the task unfolded, a trial was randomly selected (without replacement) from the bank of 100 trials. Each trial began with with a fixation cross printed in Trebuchet 28 MS font and flashed at the center of the computer screen for 200 ms. Following the fixation cross, the subliminal prime appeared for 25 milliseconds. The prime was always negative for the character selected to be subliminally negative, and was always positive for the character selected to be subliminally positive. The negative subliminal primes were: ugly, war, hurt, stink, corpse, death, hell, pain, spider, and trash. The positive subliminal primes were: flower, friend, gift, happy, kiss, smile, puppy, party, pretty, and kitten. The prime was displayed in 20pt Trebuchet MS font, and was

forward and backward masked by the nonsense word "zxcvbnm," also displayed in Trebuchet 20 MS font for 25 ms. At this point, the display appeared which included the character's picture and a potential behavioral statement. In particular, the behavioral statement appeared in Trebuchet MS 20pt font located where the prime had been. Simultaneously, the character's picture appeared centered above the behavioral statement at a size of 250x270 pixels. The picture-plus-behavioral-statement display appeared until the key d or k was pressed, reflecting the participant's choice about whether the statement was characteristic of the character or not. At that point, the appropriate feedback (either "The behavioral statement WAS characteristic" or "The behavioral statement WAS NOT characteristic") was displayed for 3000 ms. Following Rydell and McConnell (2006), each character was shown with 25 positive behavioral statements and 25 negative behavioral statements, and for each character, the positive behavioral statements were always described as characteristic, and the negative behavioral statements were always described as not characteristic, in an attempt to induce positive explicit attitudes to both characters. Following the feedback, the next trial began.

Since the hypothesis predicts an influence of the subliminal training condition on the dynamics of constructing an explicit evaluation, after training, the procedure of Spivey, Knoblich & Grosjean (2005) was followed, and participants took a mouse-tracked explicit attitude task. Participants were asked to report whether they liked or disliked various stimuli. Trials began with two seconds for participants to view the evaluative response options (LIKE) and (DISLIKE) randomly assigned to upper corners of the screen. Participants then clicked on a small box

at the bottom of the screen to reveal a stimulus word and dragged the mouse toward their selected evaluative response. The 20 stimulus pictures included the target pictures, pictures of Bob and Joe, as well as 9 positively valenced distractor pictures (e.g. John Kennedy) and 9 negatively valenced distractor pictures. These 20 stimulus pictures were presented in four blocks with randomly assigned order. The four stimulus repetitions were averaged together to yield a single measurement per participant for all statistical analyses.

Mouse-cursor movements were recorded from 25 Cornell University undergraduates. For each study, we analyzed hand-movement trajectories only from the participants who reported liking both Joe and Bob on all four stimulus repetitions. 12 subjects chose dislike on every trial for at least one picture, and were therefore excluded.

3.2.2 Pilot Study: Data Analysis Methods

Because the study intended to capture a potentially subtle influence on processing dynamics, we wanted to devise analytical measures which reduce measurement error as much as possible. In many studies which apply Spivey et al. (2005)'s mouse-tracking layout for two-alternative forced choice cognitive tasks, the researchers quantify the spatial deviations between trajectories in a fairly crude manner, measured the maximum deviation of a given point-to-point hand movement from a hypothetical straight line connecting the stimulus box to the selected

response option. However, maxima are known to be noisy statistical measures, and they reduce the complexities of a spatially selected trajectory with many points into its coordinates at a single privileged point. Thus, to obtain a more sensitive measure of spatial deviation, we wanted to develop a measure of spatial deflection that would capture information about the trajectory as a whole.

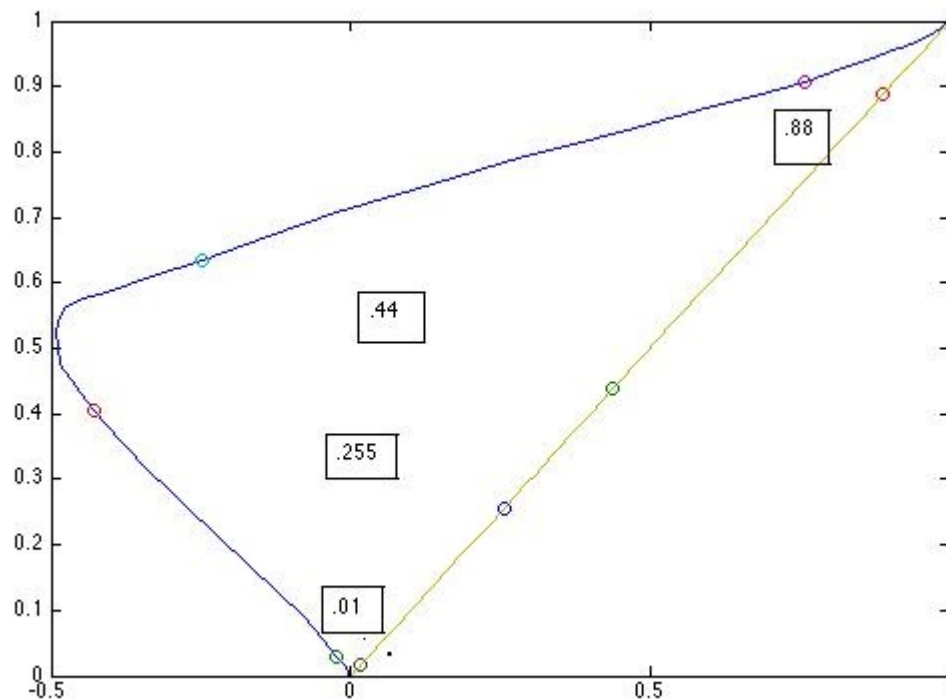


Figure 3.2: A technique for measuring the spatial deflection of hand-movement trajectories

As our computerized recording of hand-movement trajectories samples hand-

location at a frequency of 33 Hz, every obtained hand-movement trajectory is described by a discrete sample of points. We can find the Euclidean distance between any two successive points in the hand-movement trajectory for which we have data using the standard formula $d_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$, where $i = 1, \dots, i_{max}$, where i_{max} is the index of the final sample for the trajectory. Then, the total distance taken in a trajectory is given by $D = \sum_{i=1}^{i=100} d_i$. Moreover, the total distance traveled so far, D_i is a partial sum, $D_i = \sum_{j=1}^i d_j$. For example, $D_3 = d_1 + d_2 + d_3$. Finally, the percentage distance traveled is $P_i = D_i/D$. Using this method, we can simply compare each point in the hand-movement trajectory to the corresponding point in the hypothetical straight line based on matching P_i 's – percentage distance traveled. In order to do this, we first interpolate 100 distance-normalized points along each trajectory (so that this process gives us the interpolated x,y coordinates when the person has moved 1% of its total spatial distance, 2% of its total spatial distance, and so forth).

To illustrate the main idea of the metric, see Figure 3.2. In that figure, we see matching points in the trajectory for 1 percent distance traveled, 25 percent distance traveled, 44 percent distance traveled, and 88 percent distance traveled. We can make these kinds of pairings for all the points in our hand-movement trajectory. Then, we simply ask: what is the average horizontal distance between each of these points? This measure might be called "mean horizontal deflection." The horizontal deflection measure focuses on the x-coordinate, which provides the dimension of competition, rather than the y-coordinate, which is spatially or-

thogonal to the axis of competition. The horizontal deflection measure has the advantage that it uses the shape of the whole trajectory, rather than just a single outstanding point of it. We note that the spatial deflection measure distinguishes between trajectories in previous data sets as does the original maximum deviation measure. For example, in the Wojnowicz et al. 2009, Study 3, this spatial deflection measure corroborated the conclusion that when participants reported liking African Americans, their hands curved more towards the dislike response than when these same participants reported liking Caucasian Americans , $t(62)=2.56$, $p=.01$.

3.2.3 Pilot Study: Results and Discussion

In Figure 3.3, we see mean mouse movement trajectories for 13 participants. While there were not enough participants to warrant statistical analyses on trajectory properties, the trajectories do show deviation in the predicted directions.

Moreover, the table below provides distribution of dislikes for the 12 people who chose dislike every time for at least one picture,

Disliked the Subliminally Bad Guy:	8
Disliked the Subliminally Good Guy:	3
Disliked Both Guys:	1

Thus, although the small sample size prohibits statistical analyses, the pilot

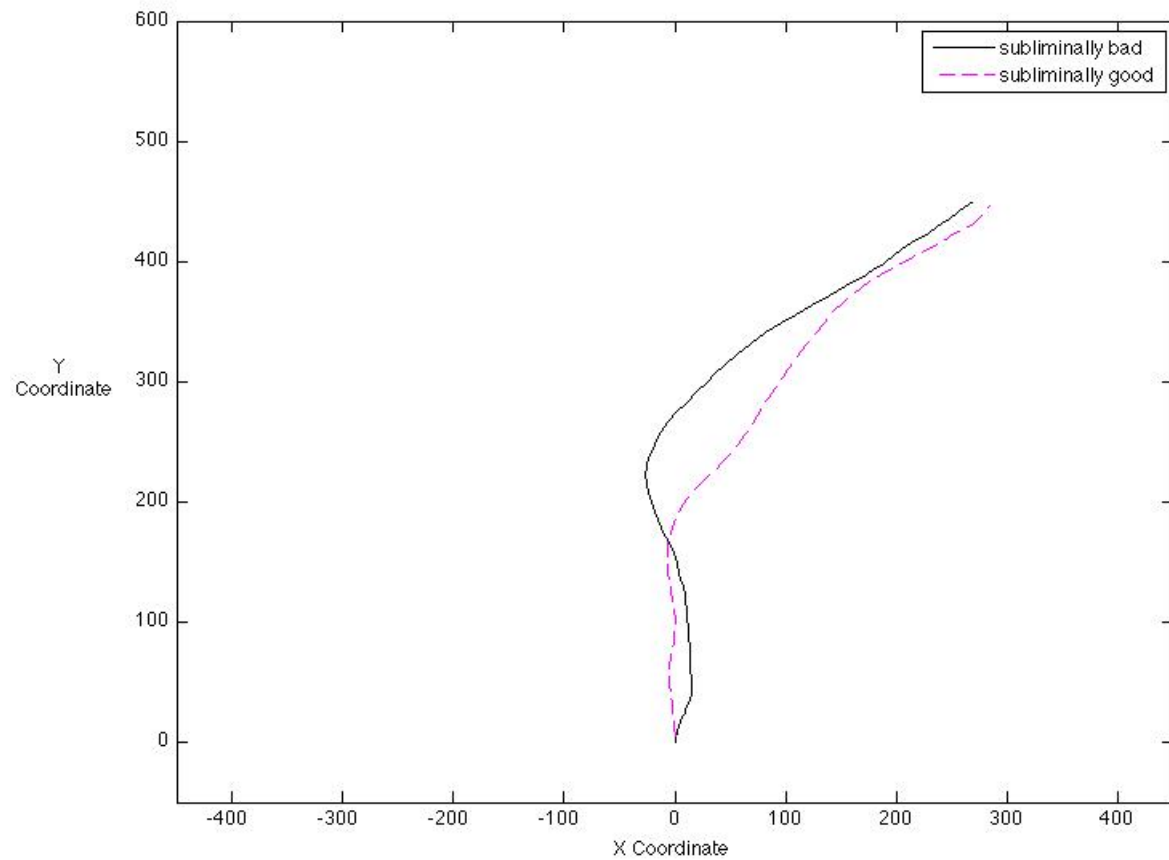


Figure 3.3: *Pilot Experiment*: Mean hand-movement trajectories towards the like decision for negatively conditioned and positively conditioned stimuli

study provides suggestive evidence that the subliminal conditioning may possibly influence the real-time dynamics of explicit decision-making, sometimes sufficiently strongly to actually push the participant fully towards a "dislike" decision.

3.3 Study 4

3.3.1 Study 4: Methods

In Study 4, we replicated the pilot finding in a larger study allowing for inferential statistics. In order to increase our effective sample size, we desired to decrease the very high percentage of pilot study exclusions (48%) caused by people consistently choosing "dislike." Thus, in this study, we modified the original methodology in the following ways:

- * In the instructions, we reminded participants to report their evaluations based on the characters' personalities, behaviors, and perspectives, rather than their appearances.

- * During pilot study debriefings, many participants stated that they selected dislike because the guy seemed too "fake" or cheesy. Therefore, we made the positive statements true 92% of the time, and the negative statements false 92% of the time, so that each stimulus person performed bad behaviors on four trials out of 25 (and likewise failed to perform good behaviors on four trials out of 25). Fur-

thermore, these behaviors were softened to be more realistic, such that statements such as "Bob always smiles at his colleagues every day" were rewritten as "Bob usually smiles at his colleagues."

* We made the distractor exemplars more extreme (e.g. Hitler), and thereby changed the response alternatives to "like" vs. "hate," assuming that participants would be unlikely to report "hating" Bob or Joe.

3.3.2 Study 4: Results and Discussion

The study had $n=93$ participants. Like the other mouse-tracking studies in this dissertation, we first exclude the participants ($n=28$) who failed to click like to both crucial stimuli on all trials. The mouse-movement plots for participants ($n=65$) who reporting liking the two characters on all four trials are shown in Figure ??.

We found greater spatial deviation towards "HATE" for the character who was negatively subliminally primed, $t(64)=2.23$, $p=.029$.

The table below displays the choices of the 28 subjects who selected "hate" for one of the men at least once.

	Subliminally Bad Man	Subliminally Good Man
Trial 1:	18	9
Trial 2:	17	7
Trial 3:	17	6
Trial 4:	17	4

Thus, this study suggests that subliminal conditioning influenced the real-time

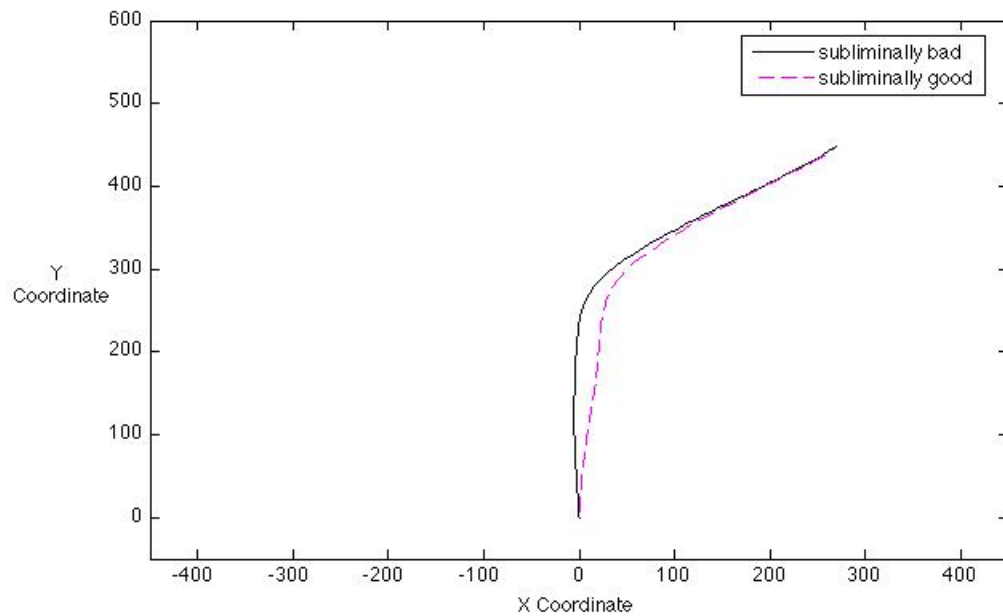


Figure 3.4: *Study 4*: Mean hand-movement trajectories towards the like decision for negatively conditioned and positively conditioned stimuli

dynamics of explicit decision-making, sometimes sufficiently strongly to actually push the participant fully towards a "dislike" decision.

3.4 Study 5

3.4.1 Study 5: Methods

In Study 5, we crafted the symmetric reversal of Study 4. That is, over the course of the 100 learning trials, we attempted to create negative explicit attitudes. Therefore, this study followed the procedure of Study 4 precisely, with the exception that the behavioral information described the two men as performing the *negative* behavior on 92% of all trials and as failing to perform the *positive* behavior on 92% of trials. As in Study 4, positive subliminal primes were flashed for 25 ms before all pictures of one randomly selected character, and negative subliminal primes were flashed for 25 ms before all pictures of the other character. The same explicit attitude mouse-tracking task was provided for this study as for Study 4, with the exception that for this symmetric reversal study, we presented the response options, "LOVE" and "DISLIKE"

3.4.2 Study 5: Results and Discussion

The study had $n=71$ participants. Like the other mouse-tracking studies in this dissertation, we first exclude the participants ($n=7$) who failed to click dislike to both crucial stimuli on all trials. The mean hand-movement trajectory plots for participants ($n=63$) who reported disliking the two men on all four trials are

shown in Figure 3.5.

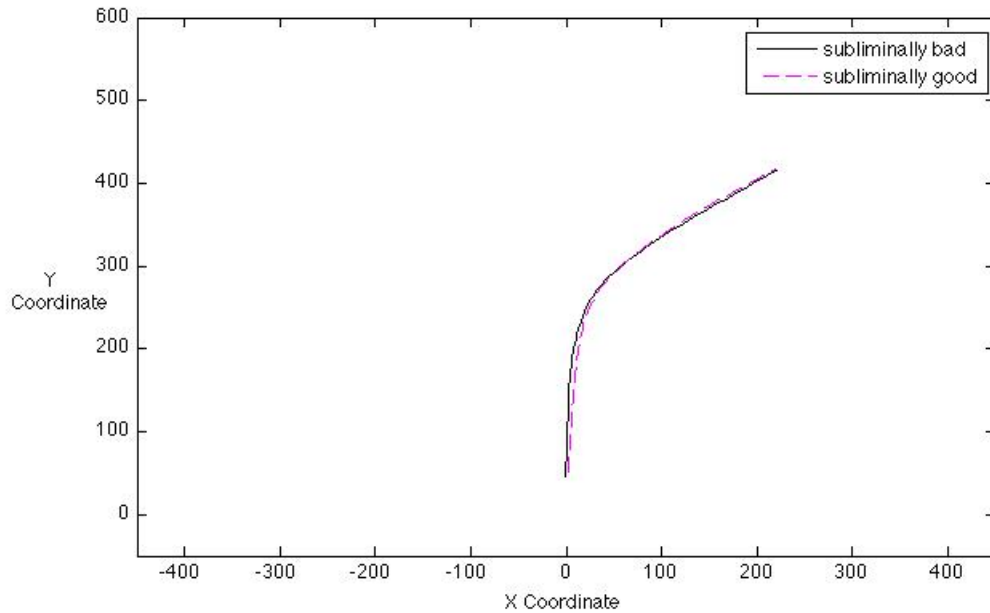


Figure 3.5: *Study 5*: Mean hand-movement trajectories towards the dislike decision for negatively conditioned and positively conditioned stimuli

This study provides no evidence to support the hypothesis that subliminal conditioning influences the real-time dynamics of explicit evaluation. One possible alternative explanation for the failure of this study is negativity dominance (Fiske 1980; Rozin & Royzman 2001; Baumeister, Bratslavsky, Finkenauer & Vohs 2001). Negativity dominance, a robust effect in the social psychological literature, refers to the fact that people pay more attention and give more weight to negative information than positive information. From a psychological standpoint, then, this study went beyond a mere flip. Since each person did the negative behav-

ior 22/25 times (and failed to do the positive behavior 22/25 times), there simply may not have been a strong enough basis for liking these people, let alone loving them.

3.5 Study 6

3.5.1 Study 6: Methods

In Study 6, we investigated the real-time influence of subliminal conditioning upon more stable attitudes. In the training task, we exposed people to sixteen concepts which people generally have strong positive feelings about, like "parties" and "summer." However, during the training task, we randomly selected half (i.e, eight) of these concepts to be preceded by negative subliminal primes.

In particular, the evaluative training task conformed to the following setup:

- 1) Inter-stimulus-interval (2000 ms)
- 2) + (centered, Trebuchet MS, 28 pt., 1000 ms)
- 3) Prime stimulus (Trebuchet MS, 28 pt., 25 ms)
- 4) Blank screen for 150 ms
- 5) Target stimulus (Trebuchet MS, 28 pt., ends on d-key or k-key)

As before, the experiment was displayed on a computer with a screen refresh rate of 85 Hz and a screen resolution of 1024x768 pixels.

Following the evaluative training, we asked participants to report whether they liked or disliked these concepts, and we mouse-tracked their responses. We wanted to compare the trajectories for the negative subliminally primed concepts to the trajectories for the positive subliminally primed concepts. Each concept was presented twice in the mouse-tracking measure, so the mean conditional trajectory for each subject included a maximum of 16 trials.

3.5.2 Study 6: Results and Discussion

The study had $n=30$ participants. Due to the relatively large number of trials (16) we recorded from each subject for each condition, we analyzed data from all subjects, discarding any trials in which "like" was the non-selected option. Figure 3.6 shows the results.

There were not significant differences in spatial deviation (and if anything, a puzzling trend in the opposite direction). However, interestingly, we found that participants' hands were dragging in their movements towards like for the concepts which received negative subliminal primes. In particular, we measured velocity using the Spivey et al. (2005) method of computing 101 t-tests over the time-normalized trajectory, and requiring at least six consecutive statistically different comparisons to conclude a statistically significant difference overall. This analysis revealed that the negative subliminal priming condition had significantly smaller x-coordinate locations from timesteps 78 to 95 (p-values: 0.0347 0.0271

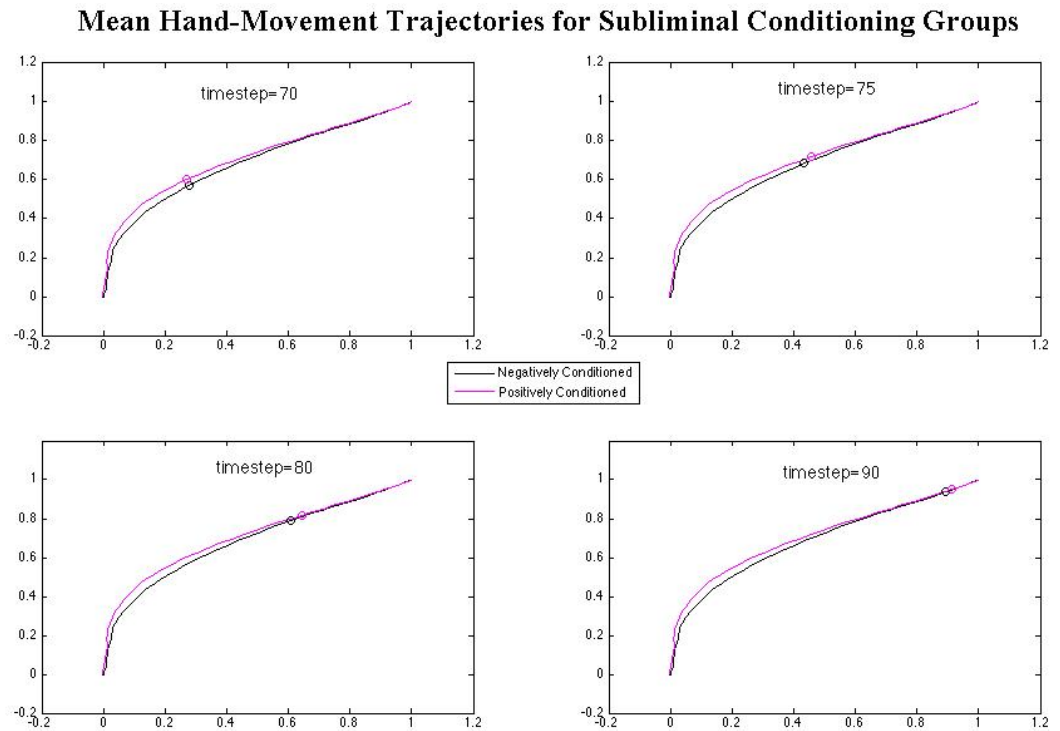


Figure 3.6: *Study 6*: Mean hand-movement trajectories towards the like decision for negatively conditioned and positively conditioned stimuli

0.0288 0.0291 0.0324 0.0281 0.0202 0.0131 0.0106 0.0091 0.0104 0.011 0.0142 0.0187 0.0261 0.03 0.0328 0.0459). This slowness covers much of the second half of their hand-movement trajectories (from about $x=0.5$ to $x=1$), as revealed in Figure ??

It is unclear why this study produced results in the domain of velocity, rather than curvature. However, this observation that the negative subliminal primes significantly dampened hand-movement velocity for approaching a "like" evalu-

ation in a spatialized decision-making layout, even for such overwhelmingly positive concepts as "sunshine" and "parties," demonstrates the power and influence of subliminal primes on the dynamics of explicit evaluation.

3.6 Conclusions

In sum, potentially contrary to many dualistic formulations whereby explicit and implicit attitudes derive from independent sources, we find possible dependencies in more sensitive behavioral measures. In particular, we find preliminary evidence that the real-time dynamical formation of an explicit attitude may be moderated by subliminal evaluative conditioning. We found that negative subliminal conditioning may be strong enough to cause feelings of hatred towards a person who performed good behaviors 92% of the time, and to cause feelings of dislike towards concepts like "sunshine" and "parties." Moreover, the evidence suggests that these temporary negative inclinations sometimes become strong enough to overpower verbal information, provoking a negative deliberative judgment. These findings, if corroborated by further research, would have potential implications for evaluative decision making in everyday life, to the extent that environmental forces (or untrustworthy manipulative agents, such as advertisers or politicians) can create classical conditioning procedures that the participant does not notice or recognize. Moreover, these findings (if corroborated by further evidence) would suggest that if people are indeed of two minds, then these minds are

not independent channels, but immediately interactive, with verbal information and subliminal information interacting at a time scale much faster than a single second.

CHAPTER 4

A DYNAMICAL SYSTEMS APPROACH TO PSYCHOLOGICAL CONTROL

4.1 Introduction: The interactive control problem

4.1.1 Psychological control is required to override implicit biases

In Chapters 2 and 3 of the dissertation, we argued that social judgment and decision-making can be described as a dynamical process of resolving conflict between (explicit or implicit) informational sources. Here we argue that the focus on conflict locates questions about the formation of an explicit social judgment firmly in the realm of psychological control. Psychological control can be defined as the ability to select thoughts and actions in the service of internal goals (Koechlin & Summerfield 2007). The main point is that, if an individual wants to achieve transiently desired goals, then he cannot respond automatically to stimuli; rather, his mind must strategically maintain certain representations, views, beliefs, and feelings, and strategically inhibit others.

What does this notion of psychological control have to do with the automaticity revolution in social psychology? If a white person with institutional power (a middle school teacher, a professional executive making hiring decisions, a city court judge, a police officer) is an aversive racist, then he may be liberal, educated, and strongly committed to an egalitarian value system, but he could still harbor

unacknowledged negative feelings and beliefs towards black people. To prevent racial bias, this person must therefore exert psychological control – strategically maintaining the representations, views, beliefs, and feelings supporting a positive view of black people, and strategically inhibiting the representations, views, beliefs, and feelings supporting a negative view of black people.

In particular, findings related to implicit racism (reviewed in Chapter 1) would suggest that the person would need to exhibit psychological control over implicit sources of information in order to express an intended social judgment. Empirical evidence supports this speculation. For example, the application of racial stereotypes (alongside other forms of heuristic judgments) increase when people are tired and presumably have depleted cognitive resources – that is, racial stereotyping is more prominent in the early morning for night people and in the late night for morning people (Bodenhausen 1990). In fact, just after white people have interacted socially with black people, they exhibit diminished performance on a Stroop task, and moreover the amount of deterioration is correlated with their implicit racism (Richeson & Shelton 2003). These behavioral results are further supported by brain imaging studies investigating the role played by various brain regions during the processing of race-based stimuli, for example the amygdala (widely believed to subserve negative emotional reactions, especially fear; Davis 1992) and the prefrontal cortex (widely believed to subserve executive control functionalities; Fuster 2008). The brain imaging studies have shown that when a person is processing race-based stimuli, their initial amygdala activation is predicted by their implicit racism scores, and that eventually their prefrontal

cortex becomes activated and suppresses the early amygdala activation (Phelps et al. 2000; Cunningham, Johnson, Raye, Gatenby, Gore & Banaji 2004).

Even outside the racial domain, psychological control is apparently required to manage the influence of implicit sources on deliberative cognitive decision-making. In the behavioral realm, people's behaviors seem to be predicted by their standards and values when control resources are high, but by their implicit attitudes when their control resources are low. For example, when control resources are high, people's eating behaviors are predicted by their dietary standards and not their implicit attitudes towards candy; yet when their control resources are low (due to emotional suppression), people's eating behaviors are predicted by their implicit attitudes towards candy, and not their dietary standards (Hoffmann, Rauch, & Gawronski, 2007). Moreover, in brain imaging studies, overriding implicit associations, which are thought to reflect people's habitual, prepotent thought patterns (e.g. insect=unpleasant; flower=pleasant), requires the activation of the prefrontal cortex (Chee, Sriram, Soon & Lee, 2000).

Thus, managing and overcoming the biases within the implicit mind seems to require psychological control. Indeed, we would expect that the mind exerted psychological control to report the explicit social judgments that it did in Chapters 2 and 3 of the dissertation. So the question immediately raised is: what is psychological control? How should we think about psychological control given the dynamic interactivity demonstrated in the first two chapters of the dissertation?

4.1.2 What is psychological control? The dualistic perspective

A major theme in contemporary psychology is that the mind can be partitioned into two separate systems. For example, in his Nobel Prize acceptance speech, the prominent economist and psychologist Daniel Kahneman popularized a pre-existing distinction between "System I" vs. "System II" (Kahneman 2002; Stanovich and West 2000). The properties accorded to these systems are listed in Figure 4.1. System I is more primitive. Its operations are fast, automatic, effortless, and difficult to modify or control. System II is obviously more sophisticated. Its operations are slow, optional, and voluntarily controlled.

Kahneman's systems are important because they capture the essence of the theoretical dualisms pervading many disparate fields of study: conscious vs. subconscious; explicit vs. implicit; controlled vs. automatic (social cognition, psychotherapy); rational vs. intuitive, deliberative vs. impulsive (judgment and decision making, behavioral economics, philosophy of mind); rule-based vs. associative; formal logical vs. connectionist ; (neural networks, cognitive psychology); cognitive vs. behaviorist (history of psychology); goal-directed vs. stimulus-directed; strategically planned vs. habitual (social neuroscience). In many cases, these dualisms are constructed as separate "systems" (planning system vs. habit system; rational system vs. intuitive system, etc.)

We bring up Kahneman's systems here because psychological theorists often think about "psychological control problems" through these dualisms. In the

Table 1. Two cognitive systems

System 1 (Intuitive)	System 2 (Reflective)
Process characteristics	
Automatic	Controlled
Effortless	Effortful
Associative	Deductive
Rapid, parallel	Slow, serial
Process opaque	Self-aware
Skilled action	Rule application
Content on which processes act	
Affective	Neutral
Causal propensities	Statistics
Concrete, specific	Abstract
Prototypes	Sets

Figure 4.1: Dual systems of cognition, from Kahneman & Frederick 2002

study of stereotyping and prejudice, people are thought to automatically intro-
ject statistical relationships (unknowingly picked up from mass media, education,
or pop culture), which may diverge from the self-avowed principles they have
constructed through careful deliberative reasoning (Devine 1989). In behavioral
economics, people are thought to display illogical behavior, such as affirming the
consequent, when simple heuristics for judgment supply an "intuition" which
can lead people astray from the more rational conclusions of effortful, controlled
deliberation (Kahneman and Frederick 2002). In social neuroscience, the brain
is believed to store automatic, habitual behaviors in certain brain structures (e.g.
the amygdala, basal ganglia, and lateral temporal cortex), but to rely upon higher
order brain structures (e.g. the prefrontal cortex, anterior cingulate cortex, and
medial temporal lobe, including the hippocampus) to monitor and override those

automatic habits (Lieberman 2006).

As suggested by Kahneman's scheme, a key feature of these dualistic theories is that the second system is posited as the controller system, and the first system is posited as the automatic system. Psychological control means that the person must exert sufficient effort and motivation so as to engage the second cognitive system and thereby supplant the potentially undesired conclusions of the first cognitive system. People fail to meet their goals when they have trouble subordinating their wild, unruly primitive system (filled with its temptations, emotions, cognitive heuristics, mere associations, and/or momentary distractions) to their more dispassionate controller system.

4.1.3 The three capacities of a psychological controller

But what exactly does Kahneman (and other dualistic theorists) mean in saying that the second system is "controlled"? We break down the capacity for control into the following three capacities which are commonly taken (in dualistic schemes) as unique capacities of the controller system:

- **Decisive selection:** By "decisive selection," we mean that dualistic theories see the executive controller system as uniquely equipped to select a decision from much larger pool of nascent candidate representations that coexist in the more primitive system. For example, it is believed that people achieve

control over unintended racist associations when the control system deliberately selects from many unintentionally activated and potentially biased knowledge structures (Kunda & Spencer 2003; see also Devine 1989, Higgins 1996), or similarly, when it selectively stamps its personal endorsement upon one of many cultural received stereotypic associations (Wittenbrink, Judd, and Park 1997; Petty, Brinol, & DeMarree 2007). The controller system is seen to exert its behavior at the "cognitive bottleneck" (Baars 1993), perhaps located in the brain's synencephalic junction (Merker 2006), where the brain selectively compresses information from a massively parallel, distributed system of highly specialized processors into a single serial stream of coherent content.

- **Strategic goal pursuit:** By "strategic goal pursuit," we mean that dualistic theories see the executive controller system as uniquely equipped to pursue distant future goals (Strack and Deutsch 2004). As Strack and Deutsch (2004) write, "While the impulsive system is driven by immediate perceptual input, the reflective system is able to abstract from the immediate input and bridge temporal gaps." A proposed function of the conscious experience associated with the control system is that it allows the person to think into the future (Baumeister & Masiocampo 2010). However, we can engage this control system only when we have sufficient motivation and effort (Fazio 1990); otherwise, we will respond to the environment reflexively.
- **Top-down flexibility:** By "top-down flexibility," we mean that dualistic theories see the controller system as uniquely equipped to flexibly switch be-

tween goals depending upon the context. From a dual systems perspective, whereas the impulsive system is relatively rigid, the controller system can adopt a strategic goal transiently (Strack & Deutsch 2004). From this perspective, a person's evaluations of black people should be a function of psychological control. Evaluations would be "stimulus-driven" by default; so barring some controlled redirection, a person's evaluation of black people would be driven by the gradual accumulation of stereotypic cultural knowledge and fear conditioning accrued by the amygdala. However, thanks to the existence of System II, participants in a psychological experiment could override their automatic racial evaluations in the service of transient goals such as pleasing the experimenter, and such an override would require the use of the prefrontal cortex and the controller system (Phelps et al. 2000; Stanley, Phelps & Banaji 2008). It is therefore the control system that allows flexible adherence to goals rather than rigid responding. It has been said that psychological control means engaging in internal, control-driven behavior, rather than simply external, stimulus-driven behavior (Monsell & Driver 2000).

4.1.4 What is psychological control? The need for an interactive theory

The dualistic approach faces a serious problem, however, in its lack of interactivity. When dual systems theories adopt a computational perspective, they posit that the controller system uses rules and the primitive system uses associations, and it is not clear how these computational forms could interact (Greenwald & Nosek 2008, Gawronski and Bodenhausen 2006). Yet in any high-performing control system, the control component and the slave component must interact with each other. A potent controller must be capable of influencing its slave system, and a responsive controller must be capable of reading the state of its slave system (Betsekas 1976). Moreover, there is a mounting empirical case supporting that Kahneman-like dual systems (the implicit, or the intuitive, or the impulsive, or the stimulus-driven vs. the explicit, or the rational, or the reflective, or the internal-driven) must interact during real-time cognitive processing (e.g. see Chapters 1, 2, and 3 of this dissertation).

Thus, a question of major importance is: what does psychological control look like within an interactive system? The dual systems perspective has elucidated what psychologists might mean by psychological control. Psychological control refers largely to three capacities – decisive selection, the pursuit of distant goals, and the flexible handling of transient goals. How could these capacities emerge from within a dynamically interactive system?

4.1.5 O'Reilly's four principles of real-time mental processing

To build a model of interactive control, we must demand that all elements of our model can be described by a single set of common operating principles. The single set of operating principles we use is the fundamental principles of biologically-plausible real-time mental processing. The four principles of biologically-plausible real-time mental processing are laid out variously across different sources (O'Reilly 1998, Spivey 2007, McClelland 1979), but they largely trace to a paper by Randall O'Reilly (1998). We will refer to them as **the four neurocomputational principles**. These four principles are described below:

- **distributed representations:** Mental representations are distributed patterns of activity across a population of neurons, i.e., distributed representations (Rumelhart & McClelland 1979). Preliminary patterns that are partially consistent with multiple interpretations of a stimulus dynamically self-organize over time into a stable pattern roughly corresponding to a unique cognitive interpretation with its accompanying behavioral response. The pattern self-organizes because synaptic connections between distributed units encode previously learned patterns of covariation, so that recurrent processing sharpens the distributed activity into a confident interpretation.
- **partial cascading:** However, a mental representation dynamically evolves in tandem with input from other brain regions as well. A brain region that projects to another brain region cannot help but transmit the constant changes in activity that it undergoes while it is processing its information.

Decades ago, it was thought that the brain, like a computer, manipulates symbols in feedforward processing stages, with neural subsystems waiting until stable completed representations have been computed before passing that information on to the next processing stage. However, it is now known that the majority of neurons in a brain region have inter-region connections, and not just intra-region connections, thus causing them to continuously cascade their evolving pattern of activation to other regions (McClelland 1979).

- **recurrent feedback:** Moreover, as a mental representation dynamically evolves, it feeds back the ongoing results of that dynamic evolution to its informational sources (its input regions). That is, brain regions that are connected to one another are typically connected bidirectionally (Douglas, Koch, Mahowald, Martin & Suarez 1995). As a result, the feed-forward cascade of information is constantly accompanied by a feedback cascade of top-down modulation of that information. This continuous recurrence of information flow means that higher-level cognitive interpretations are not completely separable from lower-level informational sources (Spivey 2007).
- **inhibitory competition:** Finally, a mental representation reaches its finalized confident form with the help of neural competition. Multiple mental representations that are partially active engage in inhibitory competition through lateral connections (O'Reilly 1998). For example, when a monkey views a visual display containing a target shape and a distractor shape (e.g. a triangle and a square), neural recordings from the inferotemporal cortex reveal that neuronal population codes for both shapes begin increasing firing from

baseline, until eventually some point is reached where the population code for the cued shape accrues greater firing at the expense of the population code for the distractor shape (Chelazzi & Miller 1993).

We note that these principles are precisely followed by the "single interactive system" models described in the introduction. Single interactive system models are an influential class of neural network models such as Normalized Recurrence (NR; Spivey 2007), Simple Recurrent Network (SRN; Elman 1990), Dynamic Field Theory (DFT; Erlhagen & Schoner 2002), and Leabra (Leabra; O'Reilly & Munkata 2000).) These single interactive system models, rather than appealing to computationally distinct formats of logical rules and network associations, attempt to explain mental functioning with a single set of operating principles. As a result, it is possible to understand how components of the model interact. Thus, we follow the single interactive systems models as a starting point for building a theory of interactive control.

4.1.6 The dynamical systems perspective on mental processing

These four neurobiological principles support a mathematical description of how mental representations dynamically form in real-time processing. The branch of mathematics is called **dynamical systems**. Dynamical systems studies how an interacting system changes over time. Dynamical systems are built out of mathematical expressions called differential equations. A differential equation is a rule

describing how some entity changes over time. For instance, a differential equation might describe the growth of a population of people given a certain birth rate, death rate. Or it might describe the upcoming firing levels of a single isolated neuron given its current firing levels and an inherent rate of decay. Now imagine a "system" composed of many possible components. Then, mathematically speaking, a dynamical system is a set of many such differential equations which are mutually dependent, or "coupled," and which describe how the state of all the components change over time. For instance, a dynamical system might describe the interactive growth of many different populations of people given birth rates, death rates, and immigration and emigration rates. Or it might describe the firing levels of many interacting neurons, since each neuron's firing rate in the future depends upon its own current firing rate as well as that of all the other neurons with which it interacts (i.e. is synaptically connected to). So a dynamical system is nothing more than a set of equations modeling how the current states of the many components of the system co-determine each other's temporal evolution. In a system with three components, a dynamical system simply takes the form

$$\dot{x} = f(x, y, z)$$

$$\dot{y} = g(x, y, z)$$

$$\dot{z} = h(x, y, z)$$

where x, y, z represent the states of some components of the system (e.g. populations levels of three countries, the firing rates of three interacting neurons), the

dot represents the temporal derivative (instantaneous change in time), and f, g, h represent arbitrary functions (i.e. rules describing the interdependencies of the components).

Dynamical systems have geometric interpretations (that is, they can be understood through spatial reasoning). Since a dynamical system in mathematics is a set of coupled differential equations and their corresponding solutions (assuming that the particular dynamical system is one for which solutions exist) , then our statement simply means there are geometric ways to interpret the behavior of these solutions. We will focus on four spatial concepts or geometric phenomena from dynamical systems theory that will help us to understand psychological control in an interactive way. For the sake of developing intuition, we will initially describe these spatial concepts not with respect to mental processing, but with respect to a poker game.

So let us imagine a poker game. There are five poker players who each put 20 dollars into the pot, and they promise to play until one player has all the money (100 dollars). We define five variables – each variable describing the amount of money a player has at a given time. Moreover, for the sake of illustration, we will assume that the particular poker game is deterministic (that is, we assume a God-like ability to determine how much money the players will have in the future based on how much they have right now and the time elapsed so far in the game). So in this setting, we can form a set of differential equations to describe how the distribution of money evolves over the course of the game. In other

words, our dynamical system simply describes the amount of money each poker player has at a given time. Of course, in a real poker game, the evolution of the dynamical system is stochastic (i.e. probabilistic – we cannot predict how much money everyone will have in the future based on how much they have right now), but we will describe this dynamical system as if it were deterministic in order to illustrate our spatial concepts.

Geometric concepts from dynamical systems theory

- **state space:** The possible states of an dynamical system with n variables. (For the poker game, $n = 5$). The geometric interpretation comes from the fact that we may assign a spatial axis to each variable in the system. Then, we can consider the state of the system as a whole as being a single point in n -dimensional space. Thus, over time, the system takes a trajectory through n -dimensional space. The state space becomes some subset of n -dimensional space which describes regions where the trajectory could possibly travel. In the example of our poker game dynamical system, the state space could most simply be described as 5-dimensional space (where each dimension describes the money held by each player). Alternatively, for a more precise description, the state space would be the four-dimensional linear subspace given by points (v, w, x, y, z) such that $v + w + x + y + z = 100$.
- **attractors:** A set of points in the n -dimensional space towards which the system will *eventually* drive itself over time if left unperturbed

by external input. The set of attractors is much smaller than the whole state space. For example, since the five poker players promise to play until one player wins, the set of attractors are given by $\{(100, 0, 0, 0, 0), (0, 100, 0, 0, 0), (0, 0, 100, 0, 0), (0, 0, 0, 100, 0), (0, 0, 0, 0, 100)\}$.

- **attractor landscapes:** A surface over the n -dimensional space which describes the pull of neighboring locations in state space on the system's current location. The common image (as in Figure 4.2) is that of a marble sitting on a rugged landscape with peaks and valleys. The marble represents the system's current location in state space, and the attractor landscape determines the movement of the marble (depending possibly on a push from external sources). When the slope is steep, the marble will move quickly. When the slope is shallow, the marble will move slowly. When there is no slope, the marble stays put (and would be settled into an attractor point). For the poker game with the rule of playing until a single player wins, the attractor landscape could take on various forms, but it must have valleys over the five points $\{(100, 0, 0, 0, 0), (0, 100, 0, 0, 0), (0, 0, 100, 0, 0), (0, 0, 0, 100, 0), (0, 0, 0, 0, 100)\}$. If player v tends to give up easily when his money is low, then the attractor landscape will have very steep descents into the locations $v = 0$, that is into regions of state space described by $(v, w, x, y, z) = (0, w, x, y, z)$. In contrast, if player v is very stubborn and begins to play extremely conservatively when his money is low, then the attractor landscape will have very shallow descents into the locations $v = 0$, that is into regions of state space described by

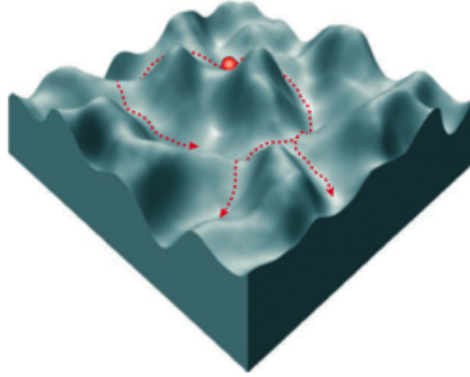


Figure 4.2: An attractor landscape.

$$(v, w, x, y, z) = (0, w, x, y, z).$$

- **projections:** We have described the state space for the poker game as living within a five-dimensional space of points (v, w, x, y, z) , and in particular being the four-dimensional linear subspace given by points (v, w, x, y, z) such that $v + w + x + y + z = 100$. The state space is currently four-dimensional, but its dimensionality could be temporarily reduced due to the imposition of temporary poker rules or temporary behavioral tendencies on the parts of the players. For example, imagine that player z is imagining that he will throw all his money out the window at the end of the game. Then although the poker game is a dynamical system taking trajectories through four-dimensional space, the true amount of money these players can expect to take home can be most accurately described by projecting these trajectories onto the three-dimensional space of points (v, w, x, y, z) such that $v + w + x + y + z = 100$ and

$z = 0$. However, consider that later on in the game, player z changes his mind. Then the system is again moving through the four-dimensional space (v, w, x, y, z) such that $v + w + x + y + z = 100$, and there is no need to project the dynamical systems' trajectories onto a smaller subspace.

Now we may describe how the geometric framework would apply to biologically plausible mental processing. Imagine a very small brain containing only three neurons, which are currently firing at rates of 20 Hz, 40 Hz, and 55 Hz. Assume that these three neurons interact with each other (that is, the firing rates of each neuron depends upon the others). Then the brain is a dynamical system, whose current location in three dimensional space is given by $(x, y, z) = (20, 40, 55)$. Of course, a real brain or brain region can have an arbitrary number, n , of components or dimensions. As the neurons interact with each other, the dynamical system takes a trajectory through that n -dimensional space. Thus, the dynamical systems perspective allows us to ask geometrical questions about mental processing, like: What are the eventual spatial destinations of the system? What are the shapes of the trajectories that the mind takes there? How does learning effect the paths that are possible?

In fact, at this point we may describe the four neurocomputational principles once again, highlighting their spatial interpretations within a dynamical systems framework. Neural systems, like the poker game system, are actually a stochastic dynamical system, but to make our primary points, it will suffice to describe neural processing as if it were a deterministic function of inputs.

The four neurocomputational principles of real-time mental processing (reinterpreted via geometric concepts from dynamical systems theory)

- *distributed representations*: Mental representations can be thought of as distinctive patterns of firing rates distributed across a set of neurons, whereby the same set of neurons can therefore encode many mental representations. Geometrically, a distributed pattern across n neurons (such as shown in Figure ??) is a point in an n -dimensional space. As these neurons interact (by sending electrical impulses through synaptic connections), the system navigates itself to certain regions of the space, "population codes", which correspond to finalized interpretations. The mind traverses into such regions of space because other patterns are unstable – they represent uncertainties in the system, and interactions between components push the state into a meaningful location (with respect to previous learning). Thus, these population codes are "attractors" of the dynamical system. Empirical neuroscience has revealed many examples of state-space population codes: for example in the hippocampus (Willis, Lever, Cacucci, Burgess, & O'Keefe 2005; Manns, Howard & Eichenbaum 2007), in the anterior cingulate cortex during a supervisory monitoring task (Lapish, Durstewitz, Chandler & Seamans 2008), in the temporal cortex during a visual object recognition task (Rolls and Tovee 1995), and in the olfactory bulb during discrimination between multiple odors (Mazor and Laurent 2005).
- *partial cascading*: Mental representations in any integrative region of the brain dynamically evolve over time, moving from indeterminate regions of

the space towards one of the population codes or attractors. Because of the principle of partial cascading, the integrative region is continually computing tentative interpretations, rather than waiting until earlier regions have fully completed their processing duties. There are many examples of this principle of gradual accumulation of information over time (e.g. Rolls & Tovee 1995; Gold & Shadlen 2000; Gold & Shadlen 2007). From a dynamical systems perspective, these higher-level regions can be described as taking state space trajectories which are evolving in tandem with the accumulation of evidence from lower-level sources of information. As the distributed pattern evolves, it becomes successively closer to the final pattern, and this dynamic evolution towards that attractor corresponds to the accrual of evidence for that interpretation.

- *inhibitory competition*: In principle, the system could beeline towards its eventual destination in a predetermined path, with the speed along that path reflecting the accumulated support for the mental interpretation. The principle of competition asserts that the system can take multiple possible trajectories to the same endpoint interpretation, depending upon the moment-to-moment evidence supporting the various candidate interpretations. In human brains, lateral inhibition (i.e. intra-layer inhibition) or normalization imposes the competition process whereby multiple candidate representations compete for limited resources (e.g. Cleland, Johnson, Leon & Linster 2007) Neural recordings have revealed many examples of competitive dynamics, whereby the system's path towards its final decision

varies by the support for various interpretations – there is competition between different visual shapes for selective attention inferior temporal cortex (Chelazzi, Miller, Duncan, & Desimone 1993), competition between multiple partially active motor behaviors in dorsal premotor cortex (Cisek & Kalaska 2005), and competition between candidate decisions in a perceptual decision-making task in LIP (Gold & Shadlen 2000). Thus, there are many (in fact, uncountably infinitely many) possible state-space trajectories towards one attractor location, and the particular pathway taken reflects the graded levels of support for alternative interpretations during the intermediate moments of real-time mental processing.

- *recurrent feedback*: The higher-level regions of the brain could continuously receive evidence from its lower-level informational sources, host a competition between candidate mental representations, and then spit out the winning representation. In that case, the system would be “feed-forward” – that is lower-level informational sources would remain sacrosanct/pristine/untouched; they would not change in response to the dynamic competition. In contrast, we see that these higher-level regions of the brain actually provide recurrent feedback to the lower-level regions of the brain, or the informational sources. That is, these higher level regions then reach back and influence the states of the incoming informational sources. Thus, this phenomenon supports the interpretation of a top-down biasing of dynamic competition (Miller and Cohen 2003; Desimone 1998). As we will see, the top-down biasing could be constructed as a projection of the

source dynamics onto lower-dimensional subspaces.

Now let us summarize the dynamical systems interpretation of real-time mental processing, considering a "zoomed-in" scale of how a neurally-implemented mental representation develops in real-time processing. More precisely, we investigate the spatial scale of firing patterns of neurons within some given brain region integrating evidence from informational sources. For instance, we could investigate the firing pattern of a population of neurons in temporal visual cortex dedicated to visual object recognition ("Is that a chair? A table? Or Michael Spivey?") based on incoming information from more primary occipital visual areas (Rolls & Tovee 1996), or we could investigate the firing pattern of a population of neurons in LIP determining visual attention ("Where should I attend to in the visual field?") based on incoming information from relevant sensory areas, such as the direction of stimulus motion in MT (Ganguli et al. 2008; Shadlen & Newsome 2001; Roitman & Shadlen 2002). The four neurocomputational properties imply that the mental representation in the integrative mental region (which for these tasks would be temporal visual cortex; or LIP) would dynamically evolve towards some stable resting population code. Initially, the incoming distributed pattern received by the integrative brain region may be partially supportive of multiple interpretations, but that distributed pattern gradually moves towards some attractor mental representation that represents a clear, finalized interpretation. These mental representation dynamics are caused by lateral inhibition, which causes some representation to win out over the others, but they are en-

hanced by the iterative cycles of information processing between the information sources and the integrator region. In particular, these iterative cycles involve partial cascading – that is, new information keeps coming in from the sources – and by recurrent feedback – that is, the source representations become increasingly constrained by the continuously evolving higher-level interpretation.

4.1.7 Goal of this chapter

The main goal of this chapter is to develop a dynamic interactive theory of control. In order to make interactions understandable, we appeal to a single set of operating principles (O'Reilly's four neurocomputational principles) so that we describe mental processing in an interactive way. Based on these operating principles, we import a dynamical systems interpretation of mental processing. Our goal is to develop an understanding of how the three psychological control capacities (selection, goal pursuit, flexibility) would look within an interactive dynamical system. We will discover that, within the dynamical systems framework, the three control capacities have clear geometric interpretations.

The appropriate understanding of psychological control is not immediate – it does not fall out of what has already been done. In neuroscience, there has been a blossoming of research into population codings and their dynamics. However, "although such processes have been investigated in some depth for perceptual and spatial domains, much less is known regarding the network dynamics that

govern higher-order cognitive processes.” (Lapish, Durstewitz, Chandler, & Seamans, 2008). In cognitive science, “it has proven difficult for the field to converge on a fully satisfying, mechanistic account of what exactly working memory is, and how it fits into a larger model of cognition” (O’Reilly, Braver, & Cohen 1999). And in social psychology, dual systems models make it unclear how the controller system’s rules and the reflexive system’s associations could interact. Thus, there is a general interest across the board in psychology in developing an interactive approach to psychological control.

One question that may arise is this: what about the single interactive system models (e.g. NR, SRN; see Chapter 1)? On first glance, it might seem that a single interactive system model could explain psychological control with all three of its capacities. The single interactive system models not only describe interactions (between units, layers, and/or brain regions), but interactions are actually part and parcel of their processing mechanisms. Moreover, the single interactive system models are neurobiologically plausible (O’Reilly 1998). Thus, single interactive system models would seem to have great potential for explaining psychological control.

However, single interactive system models have seemed able to describe only the “slow learning system” of the brain, i.e. the posterior cortex (see O’Reilly, Braver, & Cohen 1999). That is, although single interactive system models, like the posterior cortex, are capable of processing sensory stimuli and language, it turns out that they seem poorly suited for modeling cognitive control via flexible

strategic goals. We have previously mentioned two cognitive tasks that require flexible switching between strategic goals: the Stroop task and the Wisconsin Card Sorting Task (WCST). Both of these cognitive tasks require participants to flexibly toggle their cognitive processing in accordance with actively maintained strategic goals that may switch from trial to trial. In the brain, it is widely believed that good performance on these two cognitive control tasks requires the use of the basal ganglia and the prefrontal cortex to flexibly toggle between goals (Cohen, Braver & O'Reilly 1996). Computational work has arrived at the same conclusion. In particular, single interactive system networks without specialized flexible control mechanisms (which look like the basal ganglia and prefrontal cortex) seem to lack the computational capacities to excel at (1) switching between goals (toggling outputs based on flexible strategic goals) and (2) transferring knowledge between goals (e.g., Rougier et al., 2005).

Thus, to understand psychological control, we will need to turn to brain regions whose gross fundamental processing properties differ from the posterior cortex (a slow-learning, integrative brain region which the single interactive system models described above generally most closely resemble; O'Reilly, Braver & Cohen 1999). In particular, the basal ganglia has processing features enabling it to compute strategic motivational value, and the prefrontal cortex has processing features enabling it to implement flexible strategic goals (Botvinik, Niv & Barto 2009; Montague et al 2004; Botvinick 2008; Koechlin & Summerfield 2007). Thus, very recent work in the computational modeling of neurobiological systems have gone beyond the single interacting system models, incorporating multiple regions

with distinctive computational properties (e.g. Rougier et al. 2005; O'Reilly & Frank 2006; Botvinick, Niv & Barto 2009). These **multiple interacting system models** consider the single interacting system models as models of the posterior cortex, so they add additional specialized processing components reflecting the involvement of regions such as the basal ganglia and the prefrontal cortex.

However, it is critical to note that these multiple interacting "systems" are not differentiated from each other in the same manner as the dual "systems" of social psychology. In contrast to the dual systems of social psychology, all of these multiple brain systems are parallel distributed processing networks, whose cognitive processing is fundamentally characterized by interaction both inside brain regions and between brain regions. What justifies the use of the term multiple systems isn't distinct computational formats (i.e. symbolic rules versus associations) or a wall of separation between the systems (whereby communication is unclear), but rather the fact that the parallel distributed processing within these regions have distinctive specializations (in terms of neuromodulation, connectivity patterns, firing rate stability, etc.) which are functionally meaningful.

4.1.8 A note on computation vs. systems

In this dissertation, we make an important distinction between computation and systems.

As for computation, the psychological literature on control tends to describe

computation in two ways. **Dualistic computation** refers to the positing of a distinction between rule-based computation (serial processing, hard constraints, etc.) and associative computation (parallel processing, soft constraints, etc.). This kind of computation is advocated by dual systems theory. It is completely unclear how dualistic forms of computation would interact. **Interactive computation** refers to a single set of computational principles, such as O'Reilly's four neurocomputational principles, which could subserve interactions between all elements in the system described. For this paper, the term interactive computation will refer precisely to parallel distributed processing conforming to the four principles of biologically plausible real-time processing.

As for systems, the number of systems is an observer-dependent quality which depends upon the question being addressed. Mathematical models of reality inevitably abstract over existing features, but can still produce meaningful results to certain classes of questions. Even individual neurons differ in ways that are important for predicting spiking behavior based on that neuron's characteristic recovery timescales, sensitivity to subthreshold potential fluctuations, and resetting due to ionic conductances (Izhikevich 2003). For convenience, we have coined the terms "single interactive systems" and "multiple interactive systems" to distinguish between various models of social cognition. The term single interactive systems uses the term "single" because the artificial neurons in those models are largely homogenous and undifferentiated before training (although note that even networks like the SRN already differentiate between feedforward and recurrent nodes). In contrast, the term multiple interactive systems (defined above) used

the term "multiple" to reflect that fact that the parallel distributed processing in different simulated brain regions can have distinctive specializations (in terms of neuromodulation, connectivity patterns, firing rate stability, etc.) which are functionally meaningful.

In this way, we claim that it is possible for a model of social cognition to accommodate *multiple* systems, but to rely upon a *single* set of computational principles. This statement will become clearer as the argument of this chapter unfolds, but the take-home point is that the multiple systems would be defined by their specialized computational variations, but their computations would all conform to O'Reilly's single set of four biologically plausible computational principles. This "multiple systems, single computation" scheme provides an avenue for social psychologists to think about psychological control in a way that is interactive.

4.2 Selection: Dynamic contraction

In this section we argue that from a dynamical systems perspective, selection is spatial contraction. Dynamical systems offer a geometric interpretation: if each of the n components is assigned to an axis, then the state of the system can be represented as a point in that n -dimensional space. Moreover, as time passes, the system takes a trajectory through that n -dimensional space. The trajectories taken depend upon the external "inputs" to the system (i.e. where the system is initialized), as well as the internal processing dynamics of the system (the equation

governing the system's evolution). The systems may have specific stable locations, called "attractors", towards which it gravitates. (In mathematics an attractor is a region of state space that "attracts" all nearby points as time passes.) All other locations in the system are unstable, and the system will push itself away from them. Thus, a dynamical system performs "selection" by bringing itself to one of the attractors. In this way, from a dynamical systems perspective, selection is "spatial contraction": as the system continues to operate, a very large set of possible states early on is reduced to a much smaller set. We illustrate this notion with examples, and discuss the implications for selection in the mind.

4.2.1 Selection via point attractors

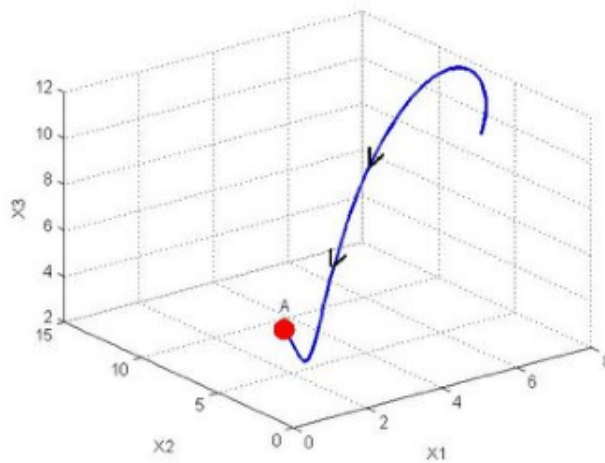


Figure 4.3: A point attractor in state space, from Chris Eliasmith

In the simplest case, a dynamical system has a single stable point – this is called a point attractor.” Consider a hayfield in Southeastern Canada where there are three perennial grasses competing for resources such as food, space, and sunlight – Kentucky bluegrass, quack grass, and timothy grass (Taylor & Aarssen, 1990). The competition between these grasses is “intransitive” ($A > B, B > C, C > A$). That is, if they were coexisting in pairs of two, the quack grass species would competitively exclude the kentucky bluegrass, the kentucky bluegrass would competitively exclude the timothy grass, and the timothy grass would competitively exclude the quack grass. But the system includes all three grasses, and this intransitive competition prevents the dominance of a competitive dominant species. Instead, there is a certain “attracting point” of species coexistence, i.e. a particular ratio of each species, to which the system will eventually move from any initial conditions, as shown in Figure 4.3. The precise location of this attracting point is set by the characteristics of the physical system – the climate, predation, parasitism, etc.

The existence of a single point attractor, besides describing the system’s eventual destination, can be considered as helping the system to resist perturbations. That is, once the system has moved into the attractor, small displacements from the attractor will relax back into the attractor. For example, consider how the brain keeps the eyes still (Seung 1996). When the eyes are still, premotor neurons (in medial vestibular nucleus (MVN) and the prepositus hypoglossi (PH)) send a constant eye position signal to the motor neurons controlling two extrocular muscles, the lateral and medial recti. The firing rates of those motor neurons in turn

are linearly related to horizontal eye position, by the equation: $v = v_0 + kE$, where v gives a vector of neural firing rates, v_0 gives the firing rates at $E=0$, and k provides the slope of influence between the eye position and the neural firing rates. The neural measures are vectors $k, v, v_0 \in \mathbb{R}^n$, and the horizontal eye position is given by the scalar $E \in \mathbb{R}$. When the system is perturbed (e.g. experimentally, if the oculomotor nerve is electrically stimulated to random v), the eye will drift horizontally from the fixed point, but the firing rates are then sucked back into their prestimulation values after a transient deflection. Thus, the eyes are held still. This feature can be represented through the bowl-shaped energy landscape (see Figure 4.4), which represents a potential function defined on the state space, in analogue with the physical idea of potential energy.

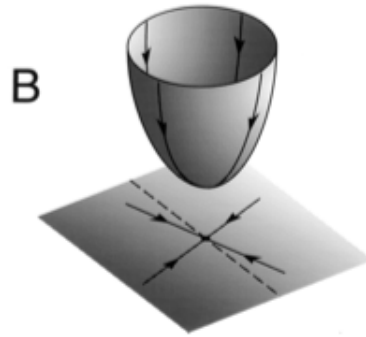


Figure 4.4: A bowl shaped attractor landscape, from Seung 1996

Although these systems are very simple, they are performing selection. They are privileging one state of the space over others – a certain stable coexistence across multiple species, a particular position for the eyes. Other examples are

commonplace: the airplane's wings are held in a consistent location despite wind intrusions, old pocket calculators compute the irrational number $\sqrt{2}$ through a dynamic procedure which brings initial state successively closer to $\sqrt{2}$, etc.

In a sense, these dynamical systems seem to be striving for a "goal." In the motivation literature, classic features of goal pursuit include: vigorous acting towards attainment of the goal, persistence in the face of obstacles, and resumption after disruption (Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trötschel 2001). For example, imagine that a young boy keeps bringing you to the ice cream shop, no matter where you are in the town, and even after you stop for a bit or distract him by showing him a cool car – in that case, you would assume that the young boy is "motivated" to eat ice cream. In other words, our very method for inferring other people's goals is by observing the above three features. But note that our simple dynamical systems share at some of these qualities: they move towards certain privileged states (they settle into attractors), and they return to these states even after disruption (resistance to perturbation).

4.2.2 Competitive selection via multistability

However, these systems don't clearly satisfy Patricia Devine (1989)'s notion of control, which we'll call the principle of "competitive selection." In her activation/application model of stereotyping, multiple associations are partially active in parallel, but the system eventually imposes a single choice. The situation of

competitive selection can be envisioned if we simply modify the earlier systems with single point attractors so that there are multiple possible attractors. In the case that competition is so fierce as to produce competitive exclusion, then there could be one attractor where the Kentucky bluegrass survives (and the other two species die); a second attractor where the quack grass survives (and the other two species die), and a third attractor where the timothy grass survives (and the other species die). Then, in early moments of the system – say, immediately after an ecological tragedy or large-scale planting endeavor– there will be simultaneous co-existence of all three species. However, over time, as the perennial grasses interact by competing for space, sunlight, and nutrients, one grass will eventually win out over the other grasses. If the system is deterministic, this eventual fate (and the intermediate states) depends solely on the initial state of the system – the system is then governed by an equation characterizing its evolution.

To illustrate, consider the Lotka-Volterra model for two competing species. The system is given by the equation:

$$\begin{aligned}\frac{dx}{dt} &= x(\epsilon_1 - \sigma_1 x - \alpha_1 y) \\ \frac{dy}{dt} &= y(\epsilon_2 - \sigma_2 y - \alpha_1 x)\end{aligned}$$

where x is the first species, y is the second species, and the other variables are parameters governing the strength of competition. We show trajectories for a particular instantiation of this system, where $\epsilon_1 = \epsilon_2 = 2$; $\sigma_1 = \sigma_2 = 2$; $\alpha_1 = \alpha_1 = 4$, as in Figure ???. Note that the system is normalized so that the number of animals are given in units (% of the system's eventual carrying capacity).

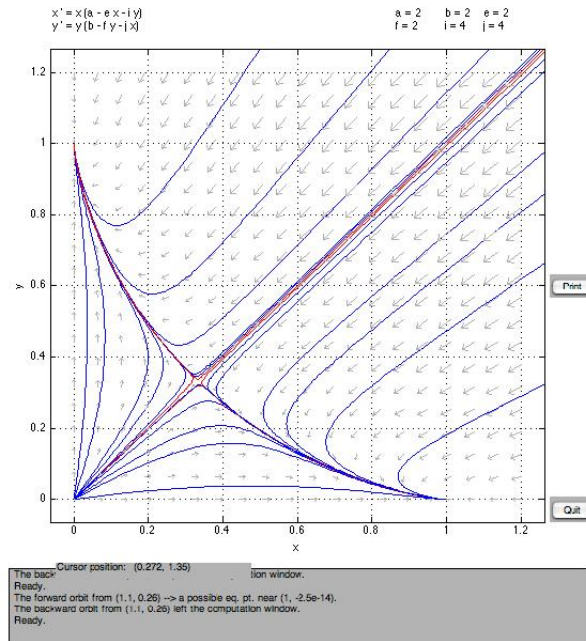


Figure 4.5: The Lotka-Volterra model for two competing species

The red lines show a "separatrix," which separates trajectories of the system with different properties. Whereas in the southeast quadrant, states flow towards the dominance of species x , in the southwest quadrant, states flow towards the dominance of species y . But different states engender different degrees of competition: the closer the system starts out to the separatrix in these quadrants, the stronger the fight by the losing species. In the northern quadrants, the system has far more inhabitants than it can hold at capacity, and so both species die off at quick rates before eventually the winning species accelerates towards the steady state carrying capacity level.

Thus, this dynamical system meets the notion of control offered by Devine

(1989) – multiple possible states (i.e. x or y) are simultaneously activated in parallel, but the system eventually selects one state over the other. But note that, in contrast to dual systems models, this selection process does not require the intervention of an external controller, but self-organizes through the interactions of the components, which are competing for limited resources. These component parts could be species competing for food (in an ecological model), or candidate representations competing for selection to drive consciousness or behavior (in a model of cognition).

4.2.3 Thesis: Selection as dynamic contraction

Dynamical systems approaches to cognition (e.g. Spivey 2007), as well as other biologically plausible models of cognition (Miller 2000; Desimone & Duncan 1995; Cisek & Kalaska 2005; Erhlagan & Schoner 2002), describe mental representations as "competing for selection." For these models, at the time scale of real-time cognitive processing, a region of the mind entertains simultaneous partial blendings of multiple interpretations, rather than having the system instantaneously teleporting from discrete interpretation to discrete interpretation (Spivey 2007). That region accumulates evidence for a given decision gradually over time, and because of the ongoing dynamic competition between multiple candidate decisions, it can take multiple possible pathways to the same eventual location. These differential pathways represent different considerations of, or different imprints from, alternative interpretations, depending on the distance to the multiple attractors at any

given time. Thus, it is said that the dynamical systems perspective emphasizes not just the symbolic, but also the "subsymbolic" (Smolensky 1995): that is, in the intermediate moments of mental processing, the mind simultaneously hosts partial coactivation of multiple candidate representations. At a mathematical level of analysis, this is no different from how in a harshly competitive ecological system, multiple species may simultaneously coexist, before one dominant species eventually wipes out the rest. This notion of competitive selection follows from the four fundamental principles for describing biologically-plausible mental processing described earlier (O'Reilly 1998), which would endow mental representations with attractor-like dynamics.

In dual systems theory, selection is a property of the controller. The associative system is often considered to be a spreading activation network (Collins & Loftus 1975), leading to the partial activation of multiple, sometimes conflicting, mental concepts. Selection occurs only when an external controller steps in to make a final selection. In contrast, in a dynamical system, "selection" has occurred without the intervention of any external power. The act of selection has not required the intervention of a separate system which operates according to distinct computational principle of logical rules. As a result, a dynamical system is said to be **self-organizing** in its selections. The initial condition or input to a competitive dynamical system generally possesses ambiguities with respect to its eventual fate – and thus its distance in state space to various attractors may be construed as tentative "interpretations" in a decision-making process (Cree, McRae, & McNorgan 1999), just as the ecological system might tentatively consider resolving into

victory for Kentucky bluegrass while heading towards a definitive victory for the timothy grass. But note that the system's own internal dynamics drive it through intermediate states into the eventual finalized decision (i.e. into an attractor), not the external intervention of an outside agent.

Thus, we have demonstrated that from a dynamical systems perspective, the first aspect of control, "selection," may spontaneously emerge from the dynamic interactions between component parts. However, this observation is just a preliminary step in solving the problem of interactive control. Selection is simply categorization plus time-dependence. But seeing a person walking down the hall, and progressively categorizing that person as the lunch lady – or hearing a blend of sounds and decomposing them into separate sources – does not fully capture the meaning of "executive control." Thus, we turn to the next property of control: strategic goal-pursuit.

4.3 Strategic goal pursuit: Attractor landscapes acknowledging distant end-states

This section is concerned with "strategic goal pursuit," which we define as: gradually manipulating the environment over the course of many actions in order to realize some distant future goals. Since this chapter of the dissertation hopes to sketch out an interactive theory of psychological control, in this section, we would

like to explain strategic goal pursuit (a particular facet of control) in an interactive way. In particular, we would like to explain how strategic goal pursuit could occur from within the scope of interactive computation, drawing upon a SINGLE set of interactive brain-like computational principles. The computational principles we will use are the four principles of biologically plausible real-time processing, as mentioned in the introduction to this chapter. We begin by discussing how strategic goal pursuit has been discussed throughout the history of psychology. As we will see, there is a long history of dualisms that prevent the interactive interpretation. We follow up by describing more contemporary, biologically-plausible interactive approach to strategic goal pursuit. We conclude by describing strategic goal pursuit from a dynamical systems perspective (and correspondingly we provide a geometric interpretation of how strategic goal pursuit is possible within an interactive computational model).

4.3.1 Dualistic perspectives throughout the history of psychology

Aristotle's perspective on goal pursuit

This mystery of strategic goal pursuit has puzzled many throughout time. When people pursue their goals, they seem to be manipulating the environment towards some final goal lying in the distant *future*. Yet a necessary condition of causality

seems to be reference to the *past*. For example, a well-known principle of statistical mechanics (Penrose 1979; see also Jaynes and Bretthorst 2003) takes it as an axiom that probabilities referring to the present time can only depend on what happened before, not what happened after. The mystery of strategic goal pursuit is this: how is it possible for a scientific model to explain how agents select actions leading them to satisfy their distant goals in the future?

As far back as Greek times, the puzzle caused problems and suggested dualistic solutions. Aristotle distinguished between two different kinds of causality – “teleological causality” – which is the desired end, or the sake for which things are done, such as walking for health – and “efficient causality” – which is the common understanding of cause, that which immediately sets a state of affairs in motion, such as the first domino knocking over the second. Because efficient causality represents the modern, scientific definition of cause as the relation of cause and effect, teleological behavior appears quite mysterious. How can apparently teleological behavior – behavior which seems purposeful and goal-like – happen in a way that is scientifically causal?

Behaviorism’s perspective on goal pursuit

Strategic goal pursuit remained mysterious for behaviorists. Behaviorism attempted to explain human behavior through its consequences: rewards and punishments. Yet one major problem faced by behaviorism is that in real life, very few actions produce an immediate slap in the face, kiss on the lips, or chocolate chip

cookie. Indeed, it is a rare life circumstance where a single behavior produces immediate primary reinforcement. In contrast, actions typically have *long-term* consequences on reinforcement. Therefore, if people want to bring about future primary reinforcement, they must learn to strategically manipulate their environment. The mystery of strategic goal pursuit is the mystery of how to explain this process.

For example, a chess player who wants to win the game must, through strategy and planning, navigate the state of the board to checkmate. At any given point, a chess player must determine the relative goodness or value of moving pieces to various possible locations on the chess board. But the goodness of any state of a chess board depends upon your partner's upcoming move, which is a stochastic random variable (i.e., you might be better at predicting it than a random number generator, but you cannot predict it with certainty). In fact, the goodness of any state depends upon events still further in the future – for example, your own move subsequent to your partner's move. But since your second next move depends upon your partner's preceding move, you cannot plan to make any particular move with any certainty. And so on and so forth, stretching out over many steps into the potentially distant future. Thus, the goodness of any state of a chess board depends upon long stretches of many stochastic behaviors, stretching into the future, each of them building upon the other. Each successive step into the future possesses an increasingly higher order of what we'll call "recursive stochasticity." This property of "recursive stochasticity" is part of what has made strategic goal pursuit seem so mysterious to models of human behavior.

Due such complications, the problem of strategic goal pursuit, like the problem of natural language, stymied behaviorism's attempt to provide a complete account of the mind. Behaviorists' primary approach to explaining purposeful, teleologically-driven behavior was through associative chain theory. They reasoned that minds could gradually create chains of associations running backwards from the goal to the starting point. For instance, a rat in a maze engages in random exploratory behavior; the rat eventually finds the goal box by chance; the sight of the goal box starts to cause small anticipatory goal responses (salivation; chewing); and eventually even these earlier stimuli become associated to anticipatory goal responses. In this way, rats eventually form a chain of associations running backwards from the goal behavior to the starting box. The proposed mechanism was attractive because it explained apparently teleological behavior through an "efficiently causal" mechanism, which therefore did not violate the laws of physics. Yet this rigid, deterministic chaining, while successful in explaining certain rat movements towards goal boxes, seemed unable to explain anything remotely as complicated and flexible as chess playing.

In its attempt to explain these more goal pursuit situations, behaviorism got stuck on the question of how to assess the "goodness" of any state (or how "satisfying" it should be to reach that state). According to Thorndike's Law of Effect, responses which lead to satisfying consequences should be "stamped in" by experience and thus occur more frequently; responses which lead to unsatisfying consequences should be "stamped out" by experience and thus occur less frequently. In situations such as the chess game, where people strategically pursue goals over

many steps, behaviorists were forced into an infinite regress: It's satisfying to capture the rook. It's satisfying to set up the board so that you can soon capture the rook... Etc. Thus, in order to explain higher-order human behavior, behaviorists often resorted to the notion of "automatic self-reinforcement" - the notion that people can reinforce themselves even when reinforcement wasn't physically present. For instance, B.F. Skinner (1961) proclaimed that a writer is reinforced by the fact that his "verbal behavior may reach over centuries or to thousands of listeners or readers at the same time." (206).

Such a stance failed, because it came across as tautological. Thorndike's Law of Effect struck many as vapid and meaningless if states can be called "satisfying" only when a person's behavior retroactively proves that it must have been so. Philosopher Daniel Dennett (1981) described automatic self-reinforcement as begging the question in order to maintain the bigger theory: "One saves the Law of Effect from persistent counter-instances by the ad hoc postulation of reinforcers and stimulus histories for which one has not the slightest grounds except the demands of the theory. For instance, one postulates curiosity drives, the reduction of which is reinforcing, in order to explain 'latent' learning, or presumes that when one exhibits an apparently novel bit of intelligent behavior, there must have been some "relevantly similar responses in one's past for which one was reinforced" (74-75). Noam Chomsky (1959) noted that in these cases, behaviorism accomplished no more than the standard "mentalese" from colloquial language: "X is reinforced by Y (stimulus, state of affairs, event, etc.)" is being used as a cover term for 'X wants Y,' 'X likes Y,' 'X wishes that Y were the case,' etc. Invoking the

term reinforcement has no explanatory force, and any idea that this paraphrase introduces any new clarity or objectivity into the description of wishing, liking, etc., is a serious delusion.” According to these critics, behaviorists were overlooking the very question at hand: *how* could a person assess the goodness associated with each possible state?

The cognitive revolution’s perspective on goal pursuit

Prompted by these failures of behaviorism, the cognitive revolution occurred, and researchers began to re-interest themselves in “hypothetical principles,” for example, the internal cognitive processes mediating between stimulus and response: attention, categorization, memory, judgment, goal pursuit, and so on. The cognitive psychologists carved out an approach to human intelligence that was antithetical to behavioristic mechanisms that simply replicated past learning. They claimed that human intelligence must go beyond the behavioristic conditioning of rats. As the Spanish physician Juan Huarte wrote in his treatise on the nature of human intelligence: “One may discern two generative powers in man, one common with the beasts and plants, and the other participating of a spiritual substance. Wit (intelligence) is a generative power. The understanding is a generative faculty.” Chomsky (1968) centered his critique of behaviorism upon this generative property. As Chomsky writes, “The normal use of human language is innovative, in the sense that much of what we say in the course of normal language use is entirely new, not a repetition of what we have heard before and not

even similar in pattern [...] to sentences or discourse that we have heard in the past."

For many cognitive psychologists at that time, a system of formal logical rules was precisely the computational scheme necessary to go beyond merely reproducing the past; that is to be able to understand and produce entirely new expressions of thought. For example, Chomsky and Miller (1958) observed that a speaker of English can understand and produce more sentences than there are seconds in a lifetime. On this basis, Chomsky and Miller argued that these sentences couldn't have been learned individually, through operant conditioning, but that instead, a far smaller set of powerful generative rules must be generating these sentences. In sum, the cognitive psychologists at the time of the cognitive revolution distinguished between two accounts of the mind. Behaviorism's learning mechanism, in responding only to past conditioning, was "merely reproductive." In contrast, cognitive psychology's account of human intelligence, thanks to its powerful rule-based structure, was freed from past conditioning and therefore "productive."

Why would strategic goal pursuit require a "rule-based system" which formally manipulates discrete symbolic representations using logical rules (see Slovic 1996)? First, recall that a rule-based system is said to declare "arbitrary variables;" a rule-based system can declare the variable "x", and reason with "x", without determining the precise value of x, and then bind concepts to these arbitrary variables later (Barsalou 1999). This computational form is believed to allow the reasoner to transcend the specific content of the material. That is, rule-

based computations allow for "universal applicability," which means the individual can exhibit equal accuracy regardless of the nature of the material – familiar or unfamiliar, tempting or untempting. This is what is meant when it is claimed that the rule-based system is a "purely logical" entity (Smith, Langston, & Nisbett 1992). For this reason, the rule-based system seems well-suited for goal pursuit. Whereas the associative system fixates on stimulus properties and their potentially unwanted motivational pulls, biases, intuitions, temptations, and superstitions (Gawronski & Bodenhausen 2006; Fritz & Strack 2004; Smith 1996; Baumeister, Bratslavska, Muraven & Tice 1998; Kahneman & Frederick 2002), the rule-based system can reason purely logically, thereby overcoming the confines of "stimulus control" and past training. Thus, rule-based system has become very basis of "executive control" in social psychology. In fact, the rule-based system is believed to underly universally applicable moral "rules," which should hold regardless of how tempting the particular situation, and which might represent the penultimate act of "executive control" or "goal pursuit."

Dual systems' perspective on goal pursuit

Following in this tradition, contemporary dual systems models encapsulate a dualistic understanding of motivation that has pervaded psychology's history. The concept of motivation is cleaved into strategic goal pursuit vs. simple conditioning. Whereas the rule-based system is capable of strategic goal pursuit, the associative system capable only of simple conditioning. For example, consider the

Reflective-Impulsive Model (Deutsch & Strack 2004), which is perhaps the dual systems model which most directly addresses motivation. This model would explain a person's reactions to chocolate M&M's in a dualistic way. On this model, an individual might automatically reach for the chocolate M&M's; on the other hand, the individual might forgo the chocolate M&M's and reach for an apple in the fridge. If the person reaches for the M&M's, the cause of behavior is located in the past; the M&M's "caused" the person to reach for them. In contrast, when people opt for the apple in the fridge, the cause of behavior is located in the distant future, when the person will achieve a very fit body.

Note how the rule-based system in contemporary dual systems theory serves the same ultimate purpose as Chomsky's production rules. Both posit the existence of a sophisticated distinct system which, due to its privileged computational form, is uniquely capable of pursuing future-located outcomes – thereby accomplishing "teleological causality." Conversely, note how the associative system in dual systems theory serves the same ultimate purpose as a behavioristic mechanism. Both are seen as possessing impoverished forms of computation, with the ultimate causes of behavior located in the past, and therefore these mechanisms are relegated to subserving "efficient causality." Thus, dual systems theory sees strategic goal pursuit as a distinct pole in a binary opposition, simply reinstating the dichotomies that run throughout the history of psychology, from Aristotle to the cognitive revolution.

4.3.2 Toward an interactive perspective on goal pursuit

Can we construct a biologically plausible, interactive model that generates convincing teleological behavior, but is scientifically causal? If behaviorists struggled because of their restrictive short-term focus on the consequences of single behaviors, then perhaps it is possible to extend behaviorism so that it handles extended sequences of behaviors. But in order to do so, we will need to formalize the notion of "automatic self-reinforcement," Skinner's notion which Chomsky found objectionable. If we could rigorously define "automatic self-reinforcement," then we could understand how when a person pursues a goal (such as trying to win a chess game), that person needn't wait for the direct reinforcement at the end of the game, but could "effectively reward" himself as he navigates effectively through the various intermediate stages of the game. That is, the player's mind could reward itself for catching the rook, or for getting into position to catch the rook, for setting up the means for getting into a position to catch the rook, etc. In short, we would want to understand how minds can compute an "internal value" which it could assign to environmental (or cortical) states, and then create internally-driven rewards for navigating to more highly valued states.

A mathematical definition of internal value

How should "internal value" be defined scientifically? This question has attracted major interest from researchers across disciplines: mathematicians (e.g. Sethian

and Vladmirsky 2001), operations researchers (e.g. French 1982), financial engineers (e.g. Glasserman 2004), computer scientists (e.g. Watkins and Dayan 1992), and neurobiologists (e.g. Botvinick, Niv & Barto 2009) – all of them loosely united under the umbrella term of dynamic programming (Bellman 1952).

Dynamic programming's first major maneuver is to transform the problem into a graph theoretical formalism. Consider examples such as people playing a chess game (and trying to select moves which maximize their chances of winning), rabbits traversing through the forest (and trying to minimize opportunities for predation), or professional movers navigating a heavy piano through a convoluted NYC apartment (and trying to minimize the work applied in joules). These scenarios can all be modeled through a graph theoretical representation where there are (a) a set of possible states (configurations of the chessboard, physical location in the forest, physical location in the apartment combined with various positioning of the piano in the hands); (b) a set of probabilities for transitioning between states (e.g. the probability of transitioning from a starting chessboard state into checkmate is 0), (c) a set of actions available at each state which influence or perhaps fully determine the transitional probabilities (e.g. the chess play selected, the chosen speed and direction of motion through the forest), and (d) a set of corresponding costs/rewards associated with these transitions (effort required to hold the piano in a certain position between two locations) and/or states (the reward of winning the chess game).

Thus, "strategic goal pursuit" can be considered in these graph theoretic terms,

where the agent is trying to minimize costs or maximize rewards over multiple steps, hoping to reach distant goals despite uncertain, reciprocal interactions with other people or the environment. The problem is to define the optimal **action policy**, which is a set of probabilities for choosing different actions given different states of the environment. In order to determine the optimal action policy, dynamic programming must construct an *internal value metric* assessing the goodness of being in any particular state. Thus, dynamic programming requires the individual or agent to engage in "automatic self-reinforcement."

Dynamic programming can define a value metric on environmental states through an expression known as the Bellman equation (Bellman 1952). This value metric compresses a great deal of complexity about the structure of a goal pursuit situation. To illustrate, let us make a simple generalization, fairly common in the neurobiological literature on motivated goal pursuit (e.g. Montague, Dayan & Sejnowski 1996), that an agent pursuing some goal will try to maximize not rewards, but returns (that is, some function of a reward sequence). A common example of a return function is an immanence-weighted sum of future primary reinforcement (Sutton & Barto 1998). For example, in discrete time, where time can be modeled as flowing in discrete stages (such as in a chess game), this return takes the form:

$$\begin{aligned}
 R_t &= \sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i} \\
 &= r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots,
 \end{aligned}$$

where the "immanence weights" γ^{i-1} are numbers between 0 and 1 and thereby simply represent a ceteris parabis preference for earlier rewards over later rewards – i.e. they are discounting terms which give successively smaller weightings to primary reinforcement value arriving successively later in time.

In this scenario, the value of a state at a given time and under some action selection policy is equal to the expected returns:

$$V_{\pi}(s_t) = E_{\pi}\{R_t | s_t = s\}$$

where $V(s_t)$ is the value of state s at time t , and π refers to the action policy chosen by the agent.

On first glance, this whole construction may seem to have accomplished little more than behaviorist theory. On the surface, the definition of "returns" R may seem to be merely allowing behaviorism's notion of rewards to inhabit different periods of time (e.g. the environment might deliver the reward at time $t+2$ instead of at time t). However, this seemingly innocuous change – and the corresponding construction of V , the internal value or goodness of being in a particular state – is a metric which compresses an enormous complex information into a single value, and which possesses qualitatively new properties. As argued in a separate document by the author of this dissertation (available upon request), the above construction of value V can be interpreted, from a dynamical systems perspective,

as the state-space distance to reward in a multistable stochastic action-dependent dynamical system. To unravel the claim, it can be shown that internal value compresses information about the probabilistic arrival into multiple end-states (multistability), combined with the time/steps needed to get to these end-states (distance), across multiple possible pathways (path stochasticity), and where an agent's own unpredictable future actions at each step in the process can change the values of all these quantities (action-dependence).

Thus, dynamical programming solves the mystery of teleological behavior. By constructing a formalized, computable metric for "internal value," dynamic programming can implement goal pursuit by simply stipulating the following principle: Actors should always select actions which maximize the "internal value" of the upcoming state. This approach has a wonderful advantage: since actors are chasing value gradients at every step, dynamical programming has transformed teleological causality into efficient causality. Thus, presuming that an actor has access to internal values, that actor can pursue goals in a way that is efficiently causal – and can therefore make "simplistic" associations between stimulus and response very intelligent indeed.

The estimation of internal value

However, the remaining question is: how can actors actually compute the values of intermediate states? So far we have merely *defined* internal value and posited that the brain, in principle, may be using this quantity to provide itself with "au-

tomatic self-reinforcement." But the problem of estimating the value of states is no trivial matter – in fact, the dynamic programming literature has devoted a great deal of research towards solving "the value estimation problem" (e.g. Bertsekas 1976). Generally dynamic programming algorithms solve this problem through a learning procedure which involves looping through a series of values relevant to a goal pursuit problem. For a simple example, in the setting of discrete temporal stages (such as a chess game) with a predetermined number of stages, K , the procedure for value estimation would require looping backwards through: a set T_k of possible future temporal stages, a set of possible states X_{k-1} in which the agent could find itself at future stage $k - 1$, a set W_{k-1} of stochastic environmental influences w_{k-1} possible at stage $k - 1$, a set of probabilities $p_{w_{k-1}}$ which are assigned to each stochastic environmental influence w_{k-1} , a set of possible actions U_{k-1} that the agent could choose based on the state x_{k-1} and the stochastic environmental influence w_{k-1} , etc. These looping procedures for value estimation are computationally intensive. In particular, they have three distinguishing properties that would, at the level of qualitative description, seem to make strategic goal pursuit require a productive rule-based system, and to be non-implementable through an interactive parallel distributed processing mechanism. In particular, the procedure would seem to require *set-theoretic quantification* and *recursion* in order to iterative looping through sets of values, and both of these capacities have been claimed to be *sin qua non* of rule-based systems (e.g. Pinker 1997, Fodor & Pylyshyn 1988; note however that recursion has become a capacity of single interactive systems since the dynamical revolution in cognitive psychology.). Moreover, the value

estimation algorithms often begin by knowing only the reinforcement properties of the end states, and the recursive loops described above would generate internal values by *backwards-computing* from the end-goal. Such an algorithm would be said to be impossible by behavioristic or associative systems (e.g. Chomsky 1969; Fodor and Pylyshyn 1988), because they would seem to require backwards propagation algorithms in neural networks, which are believed to be biologically implausible.

The above considerations suggest that strategic goal pursuit would require an entirely different mechanism or system than the constructed system of “mere associations” – thus supporting the claims of dual systems theory in social psychology. However, recent research has uncovered how internal values can be determined through a simple, local, biologically plausible procedure known as the **temporal differences method** (Sutton 1998). To illustrate, let us assume the existence of an internal “critic” whose job is to predict the value of a given state:

$$P_t \approx V_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$$

If the predictions are correct, then two successive predictions must both satisfy the following equations.

$$\begin{aligned} P_{t-1} &= r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \\ P_t &= r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \end{aligned}$$

Thus, the internal critic can exploit a convenient consistency principle which arises from the simple subtraction of power series. Namely, a correct prediction of internal value must satisfy the condition:

$$P_{t-1} = r_t + \gamma P_t$$

Thus, the critic can compare two adjacent predictions (along with the delivery of some amount of reward), and thereby improve the accuracy of its predictions. The extent to which two adjacent predictions fail to satisfy this consistency condition is the "temporal differences error":

$$\delta = r_t + \gamma P_t - P_{t-1}$$

If the temporal differences error is positive, then the earlier prediction P_{t-1} underpredicted its target $r_t + \gamma P_t$, and thus the earlier estimation of value should be increased in magnitude. And vice versa if the temporal differences error is negative. In this way, the critic can use the temporal differences error to update its predictions. It has been shown (Sutton & Barto 1998) that under fairly general conditions, this updating method will eventually cause even wildly incorrect or random initial predictions (P_t) to converge to the correct values. (V_t).

Note the import of this mechanism: The original prediction problem which required computations extending out into the (potentially infinitely) distant future has now been reduced into a consistency condition that involves only two adjacent predictions! This very fact undergirds the name "temporal differences

method” – the error which drives learning depends merely upon a difference between information available at two neighboring moments in time. As a result, there is no longer a need to worry about stochastic recursion or backwards propagation, for the same reasons as in the simple operant condition case – we are only comparing adjacent time steps. Moreover, this mechanism learns values through sampling of the environment, rather than relying upon set theoretic quantification over a known model of the world (which it doesn’t have). Finally, the mechanism is a biologically plausible bootstrapping method, meaning that, unlike the more theoretical Monte Carlo techniques, the mechanism can teach itself on-line, rather than having to wait for the final outcome to learn.

4.3.3 Brain implementation in multiple interacting systems

The basal ganglia and the actor/critic model

In the brain, a subcortical region known as the basal ganglia is believed to be important for implementing temporal differences and therefore strategic goal pursuit (Barto 1994). The basal ganglia is an interconnected set of regions, as shown in Figure 4.6, that are connected to the cortex and thalamus.

To model motivated cognition, the striatum of the basal ganglia is sometimes (e.g. Houk, Adams & Barto 1995) cleaved into an “actor” (located in the dorso-lateral striatum; denoted DLS in Figure 4.7) and a “critic” (located in the ventral

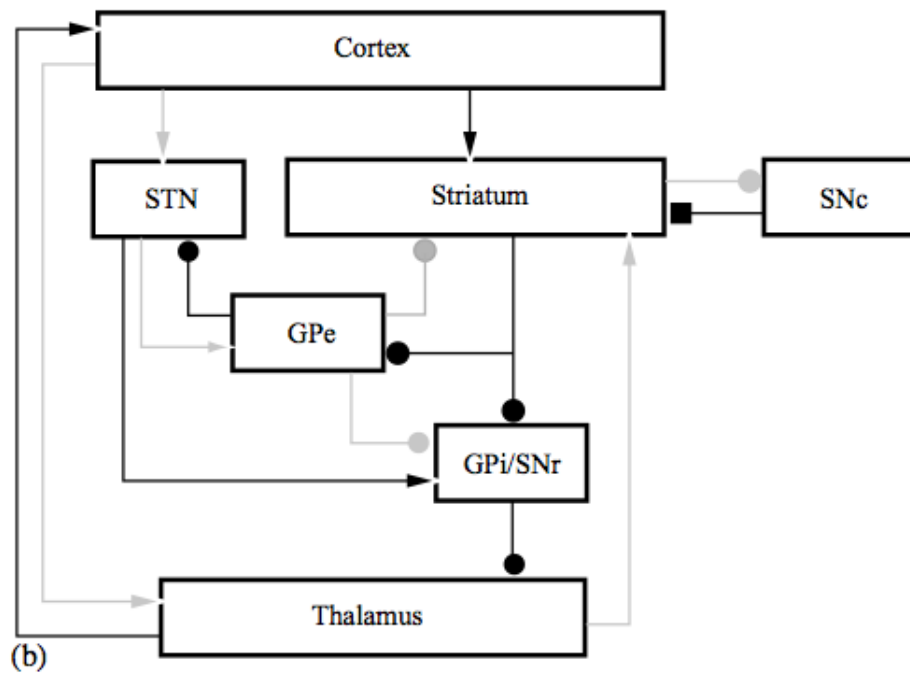


Figure 4.6: Many recurrent loops within the basal ganglia, from Bar-Gad et al. 2003

striatum; denoted VS in Figure 4.7).

The critic's job is to learn the internal value of various (cortical) states. Thus, the critic is part of the brain which can provide "automatic self-reinforcement" in the derided idea of Skinner. When the cortex has transitioned into a more highly valued state than expected, the adaptive critic should provide "effective reinforcement," even if there is no primary reinforcement directly from the environment. In order to perform this task, the adaptive critic is believed to learn value representations through the computation of the temporal differences error, which helps

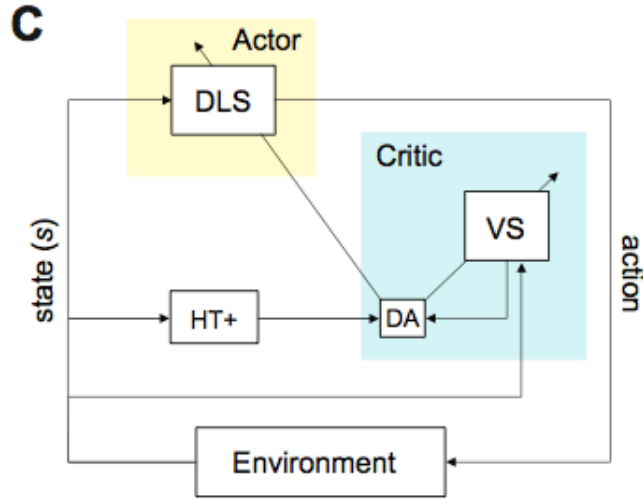


Figure 4.7: The actor/critic model of the basal ganglia, from Niv et al. 2010

it to adjust its value representations to greater accuracy. The temporal differences error is defined in equation (4.1) below:

$$\delta = r_t + \gamma P_t - P_{t-1} \quad (4.1)$$

Most, but not all, papers in the literature specify that the basal ganglia reports a temporal differences error through the firing of dopamine neurons, often in the SNc. The dopamine diffusely enters the synapses connecting sensory neurons critic neurons. Since dopamine has neural plasticity properties (Reynolds & Wickens 2002), the learning at these synapses will bolster or atrophy in accordance with the magnitude of the temporal differences error (Nakahara, Amari & Hikosaksa

2002) . In this way, the synaptic connection weights from the sensory cortical regions to the striatum's critic neurons eventually converge on storing the internal value of a state.

According to the actor/critic theories of the basal ganglia, dopamine should train the critic by reporting a temporal differences error. In particular, this means that dopamine should fire whenever the person experiences what we might call "unexpected motivational gains" – either from mispredicted rewards or unexpectedly good state transitions. As it turns out, there is evidence that dopamine fires in both situations.

Let us first consider mispredicted rewards. According to actor/critic models, if the environment delivers an expected reward, dopamine should fire above baseline levels; and if the environment fails to deliver an expected reward, dopamine should fire below baseline levels. In particular, dopaminergic firing should be a monotonic function of the extent to which the delivered reward on a given trial differs from the expected value of reward. Fiorillo et al (2003) provides supportive evidence for this hypothesis. Fiorillo et al (2003) trained monkeys in a classical conditioning procedure in which distinct visual stimuli indicated the probabilities ($P=0, 0.25, 0.5, 0.75, 1.0$) of a liquid reward being delivered after a 2-s delay. Panel "C" in Table 4.1 shows the dopaminergic responses of neurons to the delivery of reward to these various visual stimuli. The dopamine firing is a perfectly monotonic function of the extent to which a reward deviated from expectation: the lower the probability that the stimulus precedes reward, the greater the dopamin-

ergic response when the reward was delivered. Moreover, Panel "D" in Table 4.1 shows the symmetric case holds when expected rewards fail to appear. The higher the probability that the stimulus precedes reward, the greater the suppression of the dopaminergic response. These findings provide strong evidence that dopaminergic neurons are tracking reward prediction errors.

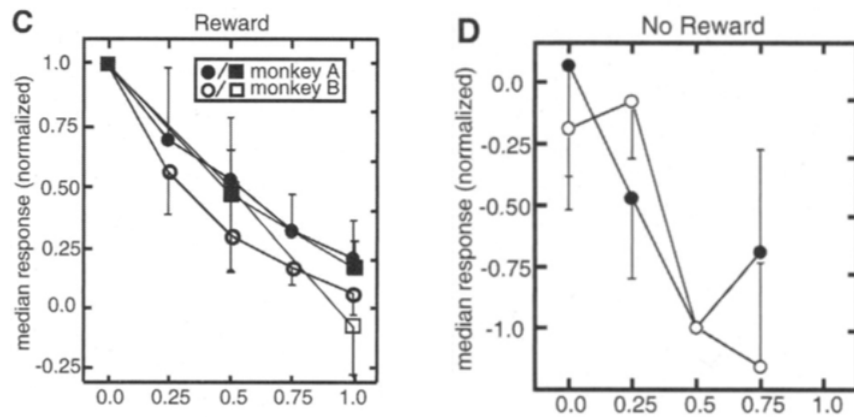


Table 4.1: Dopamine fires in response to mispredicted reward, from Fiorillo, Tobler, & Schultz(2003)

However, if dopamine is representing the temporal differences error, then its firing should predict more than just immediately delivered rewards. Rather, dopamine should be critiquing transitions between patterns of neural activity in the cortex (as argued by Montague, Dayan, & Sejnowski, 1996). Looking at equation (4.1), even if no primary reinforcement is obtained on a given trial, dopamine should fire if the cortex unexpectedly transitions into a higher valued state than it expected. In other words, dopamine should track state improvements – or rather, those that are unexpected. In the example of the chess game, if you accidentally stumble into a position where you can checkmate your opponent, then dopamine

should fire. In other words, dopamine should be functioning as Skinner's "automatic self-reinforcement."

Unfortunately, because contemporary neuroscience research has focused on reward mispredictions, there are few published paradigms providing a direct test of state improvement. However, at a gross level of analysis, it is already clear that the striatum does track state improvements. A nice example is provided by Seymour et al (2004). (However, it should be cautioned that the study was done on pain rather than rewards, and uses fMRI-BOLD rather than electrophysiological recordings of dopamine, so conclusions relevant to the actor/critic model are limited. At the current moment, papers are being generated demonstrating more directly that dopamine plays a role in state improvement (e.g. Baker & Holroyd 2009).) In the Seymour study, human subjects underwent a learning experiment in which electric shocks were delivered to their hands in response to various visual stimuli. The experimental design conformed to the methodology below:

Thus, from the beginning of a trial, there is a 50% chance of getting either cue A or cue C. Cue A is relatively highly valued (in terms of pain predictions): it has an 82% chance of transitioning into cue B determining high pain, compared to an 18% chance of transitioning into cue D determining low pain. In contrast, Cue C is relatively low valued (in terms of pain predictions): it has an 82% chance of transitioning into cue D determining low pain, compared to an 18% chance of transitioning into cue B determining high pain. Thus, Cue A is a more highly valued state than Cue C.

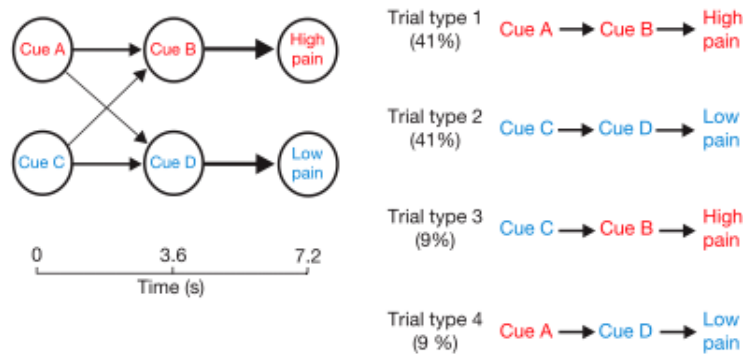


Figure 4.8: A methodology for documenting state improvement, from Seymour et al. 2004

Below are the fMRI-BOLD recordings from the ventral striatum, contrasting trial type 4 to trial type 2. What this shows is that in the first few seconds, the brain has completely unpredictably (50/50) transitioned into the more highly valued state (cue A) rather than the more weakly valued state (cue C) – therefore the right ventral striatum becomes excited from baseline. Then, the brain transitions into the same second state (cue D), a weakly valued state, and in the case when this state transition was not predicted, the right ventral striatum becomes depressed from baseline. Thus, the striatum is indeed tracking the value of stochastic state transitions in a multi-step pathway to an operant.

On the other hand, the dorsolateral striatum is known as the actor (Botvinik, Niv & Barto 2010). The actor's job is to learn action policy mappings. That is, the actor must take sensory information from the posterior cortex and make a behavioral choice. But how can the basal ganglia solve the "action selection" problem –

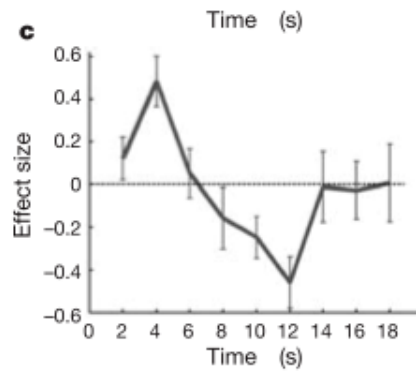


Figure 4.9: fMRI-BOLD results suggesting state improvement, from Seymour et al. 2004

that is, how can it discover which actions are expected to bring about the highly valued states with respect to its current goal? As it turns out dopamine plays a dual role. Not only dopamine train the critic, but dopamine also trains the actor. From a functional perspective, whenever the critic reports a positive temporal difference error, then the individual's action has brought him into a more highly valued state than expected, and so the individual needs to increase tendency to make that action. This functional need is subserved by the biology; dopamine is released diffusely, and when it permeates the "critic" neurons of the ventral striatum, it also permeates the "actor" neurons in the dorsolateral section of the striatum (Joel, Niv & Ruppert 2002) Thus, just as dopamine aided in the learning of state's value (by adjusting the weights between the dorsolateral striatum and the SNc), dopamine will also encourage learning of actions that are expected to maximize gains in internal value (by adjusting the weights between the ventral striatum and the posterior cortex). In this way, dopamine influences which be-

havioral choices are favored in the basal ganglia based on the potential value of upcoming environmental states.

4.3.4 Thesis: Goal pursuit as dopamine-sculpted attractor landscapes

The dynamical systems approach to mental processing becomes relevant because the actor is implemented in a real biological brain, so in principle, its computations should conform to O'Reilly's four neurocomputational axioms (which imply a dynamically evolving mental representation.) And indeed, the basal ganglia although it is subcortical rather than cortical, is well-known for its recurrent "loops-within-loops" structure (Bar-Gad, Morris, & Bergman 2003; see again Figure 4.6). Many current theories of basal ganglia function, such as its potential implementation of a multi-hypothesis sequential probability ratio test (Bogacz & Gurney 2007), documents that neurons in the output regions of the basal ganglia (GPi and SNr) implement a distributed representation, and that distributed representation dynamically evolves over time because it receives partial cascading from its informational sources, and because it forces potential action selections to compete with each other (Bar-Gad, Morris, & Bergman 2003; Bogacz & Gurney 2007). Thus, the "actor" of the basal ganglia looks very much like the integrative layer of a normalized recurrence network that was used to describe posterior cortical functionalities, despite the fact that we are now incorporating motivational

information. Thus, even in this highly "hot" motivated area (the basal ganglia is directly connected to the lateral hypothalamus believed to subserve primary rewards), the actor gradually settles upon a behavior choice over repeated cycles of activation flow.

Now let us argue that in a dynamical system, learning means sculpting the attractor landscape. In an interconnected neural system such as the brain, learning means changing the strength of synaptic connections (Bi & Poo 1998), and because the strength of synaptic weights determines the real-time dynamical interactions which will cause the system to settle into an attractor over the course of recurrent processing (see four neurocomputational principles), then changing the synaptic strengths means changing the attractor landscape of the dynamical system (see geometric interpretation of dynamical systems theory). Thus, the statement "learning sculpts the attractor landscape" is true in principle for a given neural network. Moreover, empirical research confirms that the statement is meaningful for real biological behavior. For example, it has been shown on work on bimanual coordination (Yamanishi, Kawato, & Suzuki 1980; Schoner, Zanone, & Kelso 1992) that a newly formed attractor basin can be made widened or narrowed, deepened or shallowed depending on how that newly formed basin cooperates or competes with the pre-existing structure of the entire system. In particular, the spontaneous self-organizing coordination tendencies for bimanual hand coordination have stable attractors at 0 degrees or 180 degrees. When people are trained to tap their hands at a 90 degree relative phase difference (e.g. by synchronizing with visual metronomes), the recent training competes with spontaneous tenden-

cies, and so the formed basin of attraction is shallower (signifying greater error of performance) and has a wider, less articulated shape (signifying greater variability of motion) than the deep, narrow basins of attraction that would be formed with further training at 0 degrees or 180 degrees. In fact, dynamical systems theory predicts that the entire attractor landscape changes with learning; as Kelso (1995) writes, *learning changes not just one thing, it changes the entire system*. To illustrate this, the researchers from the previous studies used a "scanning probe" over the course of many days of learning the 90 degree pattern, whereby at the beginning of each day, they tested subjects' ability to hold not just the 90 degree pattern, but all possible 12 different relative phase patterns located at discrete 15 degree steps between 0 degrees and 180 degrees. They found that learning the 90 degree pattern changes their abilities to perform the patterns at all other relative phases. As Kelso (1995) writes, "This is a far cry from the acquisition of habits and associations through repetition that have tended to dominate theories of learning in one form or another." Putting this all together, we see that *learning is the evolution of attractor landscapes*. The bimanual coordination researchers documented that different individuals have different attractor layouts at day 1, before any learning takes place. These initial layouts get morphed in idiosyncratic ways into new attractor landscapes that somehow "work" for maintaining the 90 degree pattern. Moreover, when recalling the 90 degree pattern from memory, the attractor layout snaps back in the direction of the initial landscapes.

So now we can make an argument about how strategic goal pursuit works in an interactive brain. Namely, the question is this: how should we consider

strategic goal pursuit from a dynamical systems perspective, given what we know about reinforcement learning, and given that learning sculpts the attractor landscapes of dynamical systems? To make our argument, let us return to our notion of modeling mental processing with a single interactive system model such as normalized recurrence. How should our single interactive system model get adjusted to accommodate the functionalities of the actor/critic?

Let us say the actor neurons of the striatum plays the role of the integrator region in the self-organizing normalized recurrence network, and posterior cortical regions play the role of the informational sources. The notion is reasonable, because the actor neurons in the striatum help determine the behavior choice based on posterior cortical information about the sensory environment. So in that case we have a Normalized Recurrence network between the posterior cortex and the dorsolateral striatum. Now let us ask about the role of the critic and dopamine for the Normalized Recurrence network. Recall that dopamine trains the action policy in the basal ganglia by modulating the strength of the synaptic connections between cortex and dorsolateral striatum. But in a normalized recurrence network, the strength of synaptic weights between sources and integrator governs the "partial cascading" of information during the cyclic exchanges of information between source and integrator. Moreover, since dopamine influences which actions tend to get selected over others, dopamine will end up (indirectly) influencing the "lateral competition" determining the competitive dynamics between behavioral choices.

The upshot is that, so long as dopamine firing depends upon the critic's determination of value, and that dopamine influences how the system "learns" to integrate behavioral choices out of informational sources, then we can say from a dynamical systems interpretation that the mind takes trajectories in a high-dimensional states space where value permeates the space – and the continuous trajectories are driven at every step by the implications for goal-pursuit. In other words, thanks to the work of the critic, the attractor landscapes acknowledges distant end-states. We summarize this point as saying "motivation sculpts the attractor landscape." Note that this geometric notion describes how strategic goal pursuit would work within an interactive distributed system – it does not stipulate the intervention of some computationally external duality, such as a distinct generative rule-based system, or a non-efficient kind of causality.

Setting aside the geometric interpretation, note that the dynamical systems perspective supports an interactive computational view, rather than a dualistic computational view, of goal pursuit. Contemporary theorizing in the psychological sciences (e.g. Sloman 1996; Smith & Decoster 2000; Deutsch & Strack 2004; Bodenhausen & Gawronski 2006) assumes the mind contains two different kinds of systems: an associative system which learns rote, immediate reflexive responses through simple conditioning, and a rule-based system which develop strategic plans for pursuing distant future goals. Thus, the associative system is "dumb," falling prey to simple heuristics (Gilovich, Griffin, & Kahneman 2002), whereas the rule-based system is "smart," providing the basis for rationality and higher intelligence (Andersen 1982). In this section, we have described how a

merely “associative” stimulus-response mechanism could subserve the complex, strategic pursuit of distant goals. In particular, the section focused upon computational models and neuroscientific evidence for actor-critic models of learning in the brain. In such models, people’s brains contain an internal critic which tracks the “value” of transitions between various environmental states (or their cortical representations). Whenever the person reaches a more highly valued state than expected, the critic sends out an internally manufactured dopaminergic reward signal. Using this signal, the adaptive critic can train the agent to choose behaviors which maximize the likelihood of transitioning into a more highly valued state. In this way, the mind forms stimulus-response preferences which are intelligent. That is stimulus-response associations are more than just knee-jerk reactions to immediate pleasure and pain; they can instead maximize rewards expected in the arbitrarily distant future, received with some probability after agents have guided themselves through potentially complicated stochastic pathways towards the eventual goal state.

4.4 Flexible top-down processing: The flexible selection of attractor landscapes

In the previous section on strategic goal pursuit, we described how a dynamical system could gravitate automatically towards mental representations or behavioral choices that would strategically maximize the expected value of states in the

potential distant future, even in the face of complex uncertainties. However, the value of a state can only be defined with respect to a goal. Imagine you are a college student sitting in your car at an intersection where the road to the left leads to a convenience store with cheap beer, and the road to the right leads to the campus library. Under an academic achievement goal, you must turn the right to find more highly valued environments, but under a socialization goal, you must turn to the left. The point is that internal value depends upon ultimate goals, but ultimate goals are transient and can shift wildly from moment to moment depending upon new information. Aligning one's behavior with respect to transient goals is part of what is meant by psychological control – in particular by the capacity of "flexible top-down processing." This capacity for top-down flexible toggling between goals is required to perform well on executive control tasks such as the Stroop task and the Wisconsin Card Sorting Task, as both of these tasks require transient shifts in goal pursuit in response to environmental cues.

In this section, we review recent literature arguing that single interactive systems (and the posterior cortical functionalities that they usually model) are unable to provide such top-down flexibility. We go on to review how the neurobiological structure of the prefrontal cortex has precisely the right properties to subserve top-down flexibility. Then we integrate these observations into our dynamical systems interpretation of psychological control. In the end, we arrive at a description of how to think about top-down flexibility from the perspective of interactive computation in a parallel distributed system, rather than from the perspective of dualistic computation.

4.4.1 Working memory: The information processing requirement

In this section, we will make an information processing argument that single interactive systems are insufficient for modeling control. In particular, we will argue that single interactive systems lack something called “working memory,” which is required for top-down flexibility.

Let us start by describing a cognitive task that would be difficult for the single interactive system models (NR, SRN, DFT, LEABRA). In particular, we will consider tasks related to limited-capacity visual attention (Desimone & Duncan 1995). In these tasks, participants stare at a cluttered visual array, when suddenly a target appears. The target appears very briefly and is ambiguous (e.g. it might be the letter C, but it might be the letter O). When participants can correctly provide information about this briefly flashed visual target, they receive some reward. Thus, the participant is motivated to perform well at the task. In certain versions of this task (e.g. Bisley & Goldberg, 2003, 2006), visual cues (e.g. a yellow triangle) appear, providing some information about upcoming spatial location of the target (e.g. it will be somewhere in the northwest quadrant). However, the participant cannot immediately saccade to that region of visual space; rather, the participant is required to look at a fixation cross at the center of the screen, and can only saccade when the target is flashed. Based on the structure of this task, we could say that the participant has an internally maintained goal to look at a certain region of

visual space (e.g. the northwest quadrant).

The task can be made into an executive control task by introducing stimulus-driven perturbations to the internally maintained goal. For example, while the participant is maintaining spatial attention somewhere, perhaps in the northwest quadrant, during the course of that short delay, the computer program can flash a distractor within an irrelevant region of visual space, say in the southeast quadrant. The distractor provokes a conflict between the "top-down" goal and the "bottom-up" stimulus. Executive control is required to re-establish the internally-maintained goal (to look at the northwest quadrant) in the face of the external perturbation (which automatically draws attention to the southeast quadrant; for supportive arguments about the use of executive control in these tasks, see De Fockert, Rees, Frith & Lavie 2001).

Let us now describe a simple interactive system that could perform this task. In particular, we envision a normalized recurrence network. The integrative layer would be a layer of lateral interparietal cortical neurons, as the LIP is known to provide information about spatial attention (Colby et al. 1996; Gottlieb, Kusunoki, & Goldberg 1998). The LIP layer is indeed described by distributed representations or "population codes," as the pattern of activation in LIP neurons is sufficient to determine where the person is devoting their spatial attention (Bisley & Goldberg, 2003, 2006). Moreover, LIP engages in dynamic cycles of activation exchange with lower level sources of visual featural information in the manner described in the introduction. Thus, we could construct a normalized recurrence

network which roughly captures the processing dynamics between LIP and occipital sources of visual features, and test its performance on the task.

Let us first describe when this single interactive system would succeed on the Bisley and Goldberg task. After all, the recurrent network of LIP neurons with its sensory informational sources can easily enforce a "top-down" goal, according to the structure of the attractor landscape encoded in its connectivity patterns. For example, imagine that the participant needs to fixate spatial attention at the northwest quadrant of the visual field. Then this network of LIP neurons simply needs to have an attractor at the corresponding region of neural state space (i.e. the connection weights should be constructed so that any n -dimensional state vector of LIP neuron firing rates \mathbf{x} will eventually settle into some attractive location corresponding to maintaining attention in the northwest quadrant \mathbf{x}_0). In such a situation, regardless of the current state of the network, when the distractor stimulus appears in the southeast, that stimulus will be processed by the LIP, disrupting the pattern of activation corresponding to the maintenance of spatial attention in the northwest (Bisley and Goldberg 1993, 1996). The previously stable firing pattern of the LIP neurons becomes disorganized. In other words, the distractor stimulus is an external perturbation which pushes the trajectory out of its attractor. From a more folk psychological perspective, the distracting stimulus has grabbed the participant's attention "automatically," and the participant's mind needs to exert some control in order to reinforce the internally maintained goal over the external distractor.

Who enforced this control? Nobody, if the dynamical system is wired correctly. The control is enforced automatically. If the attractor landscape is constructed to have an attractor for the northwest quadrant, then over the course of 300-400 milliseconds, the LIP will morph back into the original attractor state. The top-down goal can be enforced upon the external distractor, quite automatically, due to the simple fact that the dynamical system has an attractor set for the northwest quadrant. So there is a somewhat misleading nature to the terms that are sometimes used to describe this task – that the task requires top-down control vs. bottom-up processing, or internal control rather than stimulus-based control.

Seemingly, we are done with this section. We have already described the third functionality of control (flexible top-down processing), without any new machinery than an ordinary dynamical attractor network. In fact, the brain implemented top-down control over the stimulus, simply due to the fact that a dynamical system can perform a selection. In fact, as reviewed in the goal pursuit section of this chapter, we could do even better, extending this capacity to more complex environments (where rewards are not immediately offered, and so a person must pursue state improvement). In this case, as we have already described, the capacity for top-down control could even be made “intelligent” in the sense of guiding the system to the right states to maximize expected future rewards, simply with an adaptive critic teaching this dynamical system how to maximize the value of upcoming states.

Mommy, where do attractor landscapes come from?

However, the problem for our model is this: where does the attractor landscape *come from*? The task demands undergo dramatic changes from trial to trial: attention might be required first in the northwest, then in the northeast, then in the southeast, and then in the northwest again. And, recalling the “productive” nature of language, a person can follow any sequence of cues, even sequences it has never experienced. Somehow, the mind must be capable of implementing immediate, dramatic shifts in the attractor landscape. This functional demand requires changes to the attractor landscape that are quite different than in the Kelso bimanual coordination studies, where the adaptive critic could gradually train the attractor landscape to slowly morph into the necessary form over the course of days. In this simple visual selective attention task, the mind must immediately shift the internally-maintained goal from trial-to-trial, based on whatever cue happens to appear. This information processing demand is precisely what is meant by “top-down flexibility”

Can’t a standard single interactive system accomplish this feat? Cognitive theorizing about informational processing demands suggest that the answer is “no” (Rougier, Noelle, Braver, Cohen & O’Reilly 2005; Reynolds and O’Reilly 2009), and that these working memory tasks require additional components. To outline these components, we need to investigate the information processing demands more thoroughly.

Information Processing Demands of Flexible Top-Down Processing

- **flexible updating:** In the Bisley and Goldberg task, a fast learning system would be necessary to rapidly associate the arbitrary cue (e.g. a yellow triangle) with the transient goal to maintain attention at a particular region of visual space (e.g. the northwest quadrant). There are theoretical reasons to believe that a single interactive system which works like the slow-learning posterior cortex would not handle this demand very well. The argument traces back to classic papers on the functional incompatibilities between "fast learning" vs. "slow learning" systems (McClelland, McNaughton, & O'Reilly, 1995). For example, birds must learn species-specific songs, which is fostered by having sensitive periods early in life (when birds are around conspecifics rather than dispersing or migrating), learning templates (so they acquire only conspecific songs), and long-term retention without modification or delay (for bridging the long gaps in seasonality between breeding seasons). On the other hand, birds must recover food from stored caches, which is distributed across their home range and changes every few days. Thus, the bird must have a form of memory which rapidly updates and rapidly forgets. The memory required for these two behaviors have been argued to be conflicting and functionally incompatible. Computational work on connectionist networks provides an additional argument. Birds excel at slow gradual consolidation of memories and the extraction of shared structure. However, they famously show "catastrophic interference" on paired associate tasks (McClosky & Cohen 1989) – simple memory tasks requiring

the rapid acquisition of arbitrary associations. Thus, a "fast learning" memory system has been posited to accompany the "slow learning" memory system. This memory system is believed to reside in the hippocampus, since patients with hippocampal lesions show impairments on paired associates tasks yet preserved performance on gradually acquired skills such as reading words or tasks requiring the recovery of shared structure, such as learning of new complex stochastic grammar (Scoville & Milner, 1957; Cleere-mans, 1993). (Moreover, the hippocampus seems specialized for rapidly acquiring information through sparse, separated representations, in contrast with the distributive, overlapping representations of posterior cortex (O'Reilly, Braver, & Cohen 1999).) In the Bisley and Goldberg task, the hippocampus' ability to rapidly acquire arbitrary associations is crucial for the individual to recognize what goal should be in place (e.g. when a yellow triangle means that attention should be maintained in the northwest quadrant of visual space). But this hippocampus must be capable of projecting its rapidly learning information into a region that stores information about goals in a way that can be flexibly updated.

active maintenance: Not only must the region in consideration store goal-related information in a way that is rapidly updated, but it must actively maintain that temporarily goal-relevant information so long as it is relevant. Once the hippocampus has provided the arbitrary association to determine what goal should be in place, the mind must maintain this goal rather than let it decay. In other words, for the Bisely and Goldberg task, in

addition to the normalized recurrence like network which incorporates the LIP and sources of visual information, as well as the hippocampus to provide "fast-memory" information about temporarily associated, there must be a region devoted to active maintenance of goal-related information until the goal becomes irrelevant. Note that neither single interactive systems nor fast-learning systems can do this, for similar reasons of functional incompatibility as discussed above (O'Reilly, Braver & Cohen 1999; O'Reilly & Frank 2006)

competitive biasing Finally, the same region which actively maintains information must be capable of strategically "biasing" the competition between potential spatial regions for the limited-capacity resource of attentional focus (Desimone 1998; Desimone & Duncan 1995; Miller & Cohen 2001). If the active maintenance region is maintaining attention in the southeast, and a stimulus appears in the southeast, the bump of activation in the LIP devoting attentional resources to the southeast should persist. If the active maintenance region is maintaining attention in the southeast, and two stimuli appear in the southeast AND in the northwest, the bump of activation in the LIP corresponding to the northwest should subside, and the activation corresponding to the southeast should be maintained.

Thus, in sum, information processing requirements suggest that the mind must have a working memory system that is capable of (a) actively maintaining information over the course of the task, (b) competitively biasing the competi-

tive dynamics for limited capacity resources like attention and behavior, and (c) flexibly updating itself whenever the task demands change. Due to functional incompatibilities between the memory systems, it seems that the slow-learning, posterior-cortex like single interactive systems could not provide these functionalities.

4.4.2 Prefrontal cortex: Neurophysiological properties

Interestingly, the properties of prefrontal cortical neurons provide precisely the features necessary to subserve these information processing needs. We describe these in turn:

- **representational robustness:** First off, we know that the prefrontal cortex is specialized by a representational robustness that would subserve the active maintenance of representations (Miller 2000). Prefrontal cortical neurons can maintain activity patterns over the course of delays, and moreover, in contrast to delay-active neurons in temporal and posterior parietal cortex, they can do so even in the face of intervening sensory stimuli. (Miller, Li, & Desimone 1993; Miller, Erickson, & Desimone 1996; Constantinidis and Steinmetz 1996; Di Pellegrino and Wise, 1993).
- **dynamic gating:** Second, basal ganglia provides dynamic gating into the prefrontal cortex (O'Reilly & Frank 2006). Dopamine has a seemingly paradoxical effect on prefrontal cortical neurons: it both increases spike rates (via

enhancing persistent Na⁺ currents and reducing inactivating K⁺) and decreases spike rates (via reducing Ca²⁺ currents and depressing AMPA and NMDA components of EPSP's). As it turns out, these effects converge to stabilize prefrontal representations in the face of interfering sensory stimuli (Durstewitz, Kelc, & Gunturkun 1999). More specifically, dopamine has the triple effect of (a) strengthening the currently held prefrontal representation, (b) weakening the influence of afferent sensory information, and (c) suppressing spontaneous activity (Durstewitz, Kelc, & Gunturkun 1999; Durstewitz, Seamans, & Sejnowski 2000). Similarly, dopaminergic dips destabilize prefrontal cortical representations. Thus, recalling the arguments on goal pursuit, the basal ganglia assessment of motivational value appears to determine when the prefrontal cortex switches between actively maintained goals. Dynamic gating models of basal ganglia-prefrontal cortical loops (Hazy, Frank, and O'Reilly 2006; Montague, Hyman, & Cohen 2004) point out that, according to a temporal differences framework, dopamine fires precisely when a more valuable goal can be achieved if behavior is redirected towards that goal – and thereby causes the prefrontal cortex to adaptively let in representations which guide behavior through active maintenance. Based on these features, predominant models of executive control in computational neuroscience (Hazy, Frank, and O'Reilly 2006; Montague et al 2004) have been positing that the basal ganglia help the prefrontal cortex to “know what goals to have.”

- **recurrent hierarchical position:** Finally, the prefrontal cortex is specialized by sitting on top of a hierarchy (Fuster 1997; Fuster 2000), providing the prefrontal cortex with immediate privileged access to many disparate domains that would subserve top-down competitive biasing. For example, one researcher (Young 1993; see also Stephan, Hilgetag, Burns, O'Neill, Young, & Kotter 2000) developed a 72 by 72 entry connectivity matrix for these cortical regions, coding reciprocal connections as a 2, one-way projections as a 1, and unreported connections as a 0. In this way, each region could be assigned a coordinate in a 72-dimensional "connectivity space." Using multidimensional scaling (Shepard 1980), Young projected this space onto a two dimensional plot which optimally fits the connectivity matrix – i.e., such that the distances between the points on the plane were as close as possible to the reverse rank order of the 'proximities' between areas in the connection matrix. The results are shown in the figure. Sensory domains occupy the lower part of the figure. The visual system is on the left, the somatosensory-motor is on the lower right, and the auditory system is on the upper right. The frontolimbic domain is at the top of Figure 4.10. As Merker 2004) says, the frontal limbic region is the "common connective center of gravity of the entire cortical system."

Moreover, the dorsolateral prefrontal cortex provides rich and diverse *recurrent feedback* from its privileged position in the cortical hierarchy. The dlPFC has reciprocal cortio-cortical connections with a large array of cortical and subcortical areas posterior parietal cortex, inferior temporal cortex, superior

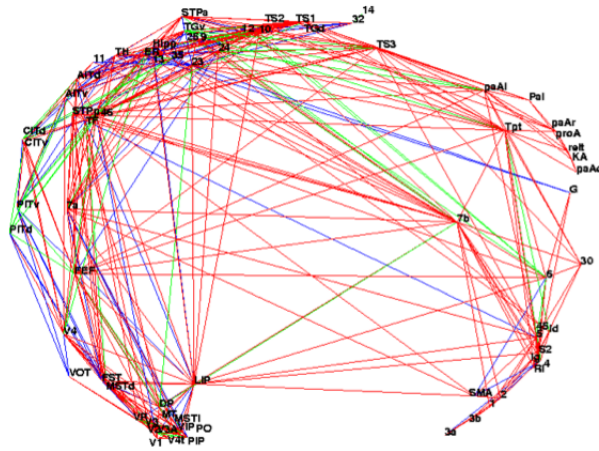


Figure 4.10: Seventy-two regions of cortex in connectivity space, from Young 1993.

temporal polysensory areas, anterior cingulate cortex, retrosplenial cortex, parahippocampal gyrus, the mediodorsal nucleus of the thalamus, and it also sends feedback to the frontal eye field, the pre-supplementary motor area, the premotor cortex, the caudate nucleus, and (indirectly, via retrosplenial area 36 and the posterior presubiculum) to the hippocampus indirectly (Funahashi 2001; Fuster 1997; Pandya & Barnes 1987). In addition, neurophysiological data from the dorsolateral prefrontal cortex suggests that this top-down recurrence modulates sensory processing, such as in visual cortical areas (Miller, Erickson, & Desimone 1996; Kastner & Ungerleider 2001), as well as a large array of other cognitive activities, such as episodic memory and retrieval (Funahashi 2001).

These considerations suggest that the prefrontal cortex could provide just the right physical mechanism for implementing working memory. The biological structure of the system matches the three informational processing requirements for a working memory system. In fact, detailed biologically plausible computational modeling has supported the notion that a specialized working memory region with prefrontal cortical properties is necessary to perform cognitive control tasks requiring top-down flexibility (e.g., Rougier et al., 2005). To substantiate this claim, let us consider two cognitive tasks that require flexible switching between strategic goals: the Stroop task and the Wisconsin Card Sorting Task (WCST). Both of these cognitive tasks require participants to flexibly toggle their cognitive processing in accordance with actively maintained strategic goals that may switch from trial to trial. The Stroop task requires a controlled override, requiring people to categorize color word stimuli according to ink color rather than word name (as is typically done; Engle 2002). The WCST requires participants to switch from trial to trial among categorizations according to the color, shape, texture, etc. of multi-feature stimuli (Miyake et al 2000). In the brain, it is widely believed that good performance on these two cognitive control tasks requires the use of the prefrontal cortex to flexibly toggle between goals (Cohen, Braver & O'Reilly 1996). The simulations by Rougier et al (2005) arrived at the same conclusion. In particular, single interactive system networks without a specialized flexible control mechanism seem to lack the computational capacities to excel at (1) switching between goals (toggling outputs based on flexible strategic goals) and (2) transferring knowledge between goals (e.g., Rougier et al., 2005). A multiple interacting systems model

– which included a basal ganglia component for learning motivational value, a prefrontal cortex component for maintaining the current goal representation, and dynamic gating between them – outperformed a single interactive system model on the both WCST and Stroop task. These simulations have supported the notion that single interactive system models seem able to describe only the “slow learning system” of the brain, i.e. the posterior cortex (see O’Reilly, Braver, & Cohen 1999). That is, although single interactive system models, like the posterior cortex, are capable of processing sensory stimuli and language, a working memory system is necessary for modeling cognitive control via flexible strategic goals.

4.4.3 Control parameters: The dynamical systems concept

So far, we have argued that top-down flexible processing requires a component that looks different than the slow-learning, posterior-cortex-like processing of the single interactive systems (e.g., NR, SRN, LEABRA, DFT). We have also identified the prefrontal cortex as the likely substrate of those functionalities. In this section, we match the features of top-down flexible processing with a concept (and corresponding geometric interpretation) from the mathematical theory of dynamical systems.

To do this, let us consider the simple dynamical system:

$$\dot{x} = r + x^2,$$

which is pictured in Figure 4.11.

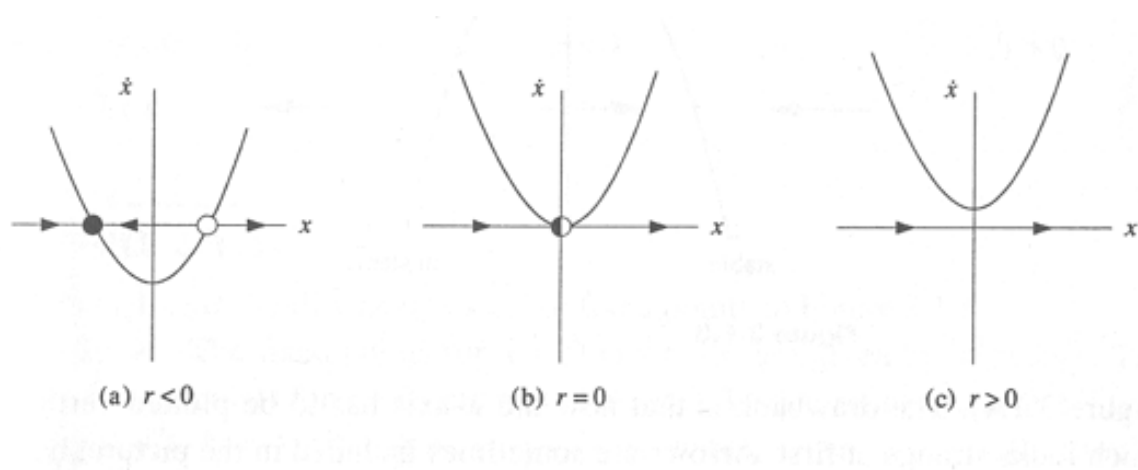


Figure 4.11: A saddle-node bifurcation, from Strogatz 1994.

Since this system only has a single dimension, it is really no more than a differential equation. That simplicity provides a nice convenience, as we can very easily understand the behavior of this system by graphing the change in x , \dot{x} , as a function of the state of x . Thus, we can read out the system's attractor landscape from the graph. The state space is the single dimension graphed on the horizontal axis, and the attractor landscape in this case is the parabola over that horizontal axis. The attractor landscape above the x -axis describes how the system will move from the point; so positive values of the parabola mean the system will increase in value from the current point, negative values mean the system will decrease in value from the current point; high negative values mean the system will quickly decrease in value from the current point; etc.

Recall the definition of an attractor, provided earlier, and note that an attractor

exists whenever perturbations to a point disappear, such that the system snaps back into the original point. In a one dimensional system such as this one, an attractor exists if and only if the graph of \dot{x} crosses the abscissa with a negative slope. So in panel (a) for instance, we see that the black dot is an attractor – neighboring points on the x -axis are “attracted” to it, in the sense that the system’s dynamics move neighboring points increasingly closer to the black dot with time.

Now to illustrate the dynamical systems interpretation of top-down flexibility, we will concern ourselves with the role of the r variable. Let us assume first that r is some positive number, $r > 0$. Then the differential equation is a parabola that is always above the abscissa, as shown in frame (c). Thus, the system will always grow without bound. For any starting initial condition, x , the value of x will always increase over time – in fact growing exponentially. Of course, because the value here accelerates off into infinity, this simple dynamical system is not perfectly realistic, at least for $r > 0$. But this very simple dynamical system could perhaps model the population growth rate: the birth rate minus death rate plus immigration rate. As an example, we can imagine that “ x ” represents the population growth rate of some insects, and “ r ” represents the amount of tree shrubs available to a population of insects.

Now, imagine reducing r slightly to r_0 , such that $0 < r_0 < r$. In this case, the parabola shifts downward slightly, meaning that the dynamics of growth to infinity have slowed down at each state x . So because of the shift in r , our population growth rate no longer shows quite as much dynamical increase. The shift

in r has shifted our attractor landscape.

Now imagine sliding the value of r down to a negative level, such that $r < 0$. Note that something very interesting has happened: there is a qualitative change in the dynamics. Suddenly, now, the net population rate only grows if it is sufficiently negative (i.e., lots of deaths and emigration). Otherwise, the net population rate will *decrease*. Reducing " r " past the threshold of $r = 0$ has had an immediate devastating effect on our population.

The point is to notice the role played by the variable r . That variable is not part of the state of the "system" – the population growth rate is given by x . Yet the variable r has an important effect on the dynamics. Whereas for $r > 0$ any initial population growth rate would increase without bound, now most population growth rates will decrease, and in fact will decrease all the way to a negative value, meaning our population will eventually die out!

In essence, if the dynamical system governs the future of the population and $r < 0$, then the population may look healthy now, but it is destined for failure. The equation has an *attractor* at the value $x = r$, and r represents a negative growth rate. Note, however, that there is one hopeful feature of this system's attractor landscape. Namely, the system can escape its bleak future if some external perturbation shifts the value of r back upwards to ranges where $r > 0$. In that case, the state of the system will grow without bound (assuming there is no meddling from external forces into the system's own inherent dynamics).

We call this variable “ r ” a control parameter, because it controls the system’s dynamics. In particular the variable “ r ” determines the attractor landscape for a given dynamical system. In our case example, when the control parameter slid past the value $r = r_c = 0$, the attractor landscape qualitatively changed. At that moment, an attractor was born (note also that a repeller was born as well, on the other side of the attractor). We call this qualitative change in the system’s dynamics, caused by the control parameter, a *bifurcation*. Moreover, we call the point r_c at which this occurs a *critical point*. However, these terms are not important for our argument; merely the more general point that the control parameter shifts around the attractor landscape of the system.

As an example, a dynamical systems model has been made by bioecologists to describe the population of a pest known as spruce budworms “ x ” living in a forest of balsam fir trees. Generally birds keep these pests in check, but sometimes the population of pests has an “outbreak” (Ludwig, Jones, & Holling, 1978; see also Strogatz 1994). Thus, bioecologists have attempted to model the population dynamics of the spruce budworm. In these models, the change in the budworm population variable (x) was modeled by a function, $f(x,r)$ dependent upon predation by birds (i.e. predation is a function of state, x) and forest variables such as the amount and proximity of trees (r). The forest variables were modeled by the control parameter (r), because trees grow and die on a much slower timescale (their lifespan in the absence of budworms is 100-150 years), than the fast timescale of the budworm population (they can increase their density fivefold within a year).

4.4.4 Thesis: Flexible top-down processing as the adaptive selection of attractor landscapes

Our thesis is that, from a dynamical systems standpoint, the prefrontal cortex exerts flexible top-down control by selecting attractor landscapes that govern the dynamics of the rest of the brain. In order to arrive at this conclusion, we must integrate the three previous subsections, exploring the three-way correspondence that exists between functional demands (i.e. the information processing requirements of working memory), neural implementation (i.e. the distinctive neurophysiological characteristics of the prefrontal cortex), and dynamical systems modeling (i.e. the role of control parameters). In particular, we show that the role played by the control parameter "r" with respect to a dynamical system is the same role played by "working memory" with respect to information processing and the same role played by "prefrontal cortex" with respect to the brain. Let us pursue this three-way correspondence with respect to general dynamical system, which can be written as

$$\dot{x} = f(r, x) \tag{4.2}$$

where x describes the state of the system in an arbitrary number of dimensions.

The three-way correspondence is laid out in Table 4.2:

Since we have already described the correspondences between the neurobi-

Table 4.2: Flexible top-down control: A three-way correspondence between model, behavior, and implementation

Facet	Information Processing Requirements for Working Memory	Properties of Prefrontal Cortical Neurons	Behavior of a Dynamical Systems Control Parameter
1	active maintenance	representational robustness	invariance to state dynamics
2	flexible gating	dopaminergic gating	superordinate dynamics
3	competitive biasing	hierarchical recurrent feedback	dependent variable to state

ology of the prefrontal cortex and the information processing requirements of working memory, in the section below we will relate the information processing requirements of working memory to the properties of a control parameter of a dynamical system.

- **Correspondence 1:** When we say that the control parameter is “invariant to state dynamics,” we mean that as the system evolves over time according to the equation (4.2). The variable r does not change; that is, r does not appear on the left hand side of the equation. Thus, even as x changes in our dynamical system, r remains constant, and it does so regardless of x . In actuality this hard-and-dry relationship is an artifact of modeling. In real-life systems, r does typically change, but the point is that it changes at a much slower timescale than the x variable. Recall that in the budworm and forest dynamical system, the forest variable were modeled by the control parameter r , because trees grow and die on a much slower timescale (their

lifespan in the absence of budworms is 100-150 years) than the fast timescale of the budworm population (they can increase their density fivefold within a year). The key point is that the control parameter r is largely invariant to the evolution of the state x .

Now we note that the notion of the control parameter r corresponds to the "active maintenance" role of a working memory system. Just as r is invariant or stable to the evolving dynamics of a system, the actively maintained goal is stable regardless of the evolving dynamics of a particular sensory representation or behavioral choice. For example, in the visual attention task, the actively maintained goal to maintain attention to the northwest should remain stable regardless of the evolving dynamics of the spatial attention system, until a new goal is provided. That is, if you should be devoting your attention to the northwest quadrant of visual space, something should keep the system's attractor landscape still, regardless of what perturbations are introduced into the system's dynamics by unpredictable visual distractors.

Thus, from a dynamical systems perspective, properly informed by systems neurobiology, the prefrontal cortex is setting up a control parameter r governing the dynamics of the spatial attention system-, and this actively maintained control parameter is what holds the attractor in desired region of visual space.

- **Correspondence 2:** Now recall the informational processing requirement of "flexible gating": namely goal-related information must govern mental processing dynamics during a certain task set, and then after some period

of dynamics under its jurisdiction, the goal-related information must switch. We observe that the control parameter r of a dynamical system provides precisely this capacity, because its dynamics are "superordinate." As discussed earlier, the parameter r does not change due to relatively fast changes in the state variable x , but it is perfectly capable of changing due to its own slower dynamics. The neurobiological observations from the previous subsection suggest that dopaminergic gating could be shunting in new values of r when task demands change.

- **Correspondence 3:** For the third correspondence, we note simply that, as shown in (4.2), the control parameter of a dynamical system is a dependent variable of the differential equation. Thus, the dynamics of the system depend upon its value. It is in that sense that the control parameter can provide a "top-down biasing" of the competitive dynamics of real-time mental processing (Desimone 1998; Miller and Cohen 2003).

These relationships suggest (especially in conjunction with the neurocomputational dynamics frequently used to model posterior cortex) that the prefrontal cortex is maintaining an r governing the dynamics of the rest of the system, x (which would correspond to areas lower in the hierarchy, including but not necessarily limited to posterior and motor cortex, and their corresponding functions of sensory categorization and action selection). Thus, from a dynamical systems perspective, the prefrontal cortex is selecting attractor landscapes. Moreover, this r is clamped into place by the midbrain dopamine systems, and thus, the firing of

dopamine causes the prefrontal cortex to maintain a specific attractor landscape over posterior cortical space. This functionality (maintaining an attractor landscape at the time-scale of working memory) corresponds to the notion that the PFC maintains "task representations" (Monsell 2003), because the given attractor landscape would presumably subserve particular dynamics for categorization and action selection relevant to the pursuit of a particular goal (that is, the attractor landscape would contain particular attractors or destinations for the dynamics, as well as a particular set of possible paths for moving to those destinations).

The notion that the brain must have a control parameter for the dynamics in order to perform well on certain tasks (namely, those requiring top-down flexible processing, such as the Bisley and Goldberg task) corresponds extremely closely with the notion of "options" (Botvinick, Niv, & Barto 2009) in hierarchical reinforcement learning. The notion of "options" arose when the machine learning community discovered that, for optimal performance in reinforcement learning tasks, agents needed to be able to select not only "simple actions," but also temporarily maintained policies (mappings from states to actions; see Barto & Mahadevan 2003). That is, the agent must be capable of selecting not only (primitive) actions, but also full-on action policies. These higher-order selections are called "options." The brain seems to have stumbled upon this very solution itself, since the neurophysiological properties of the prefrontal cortex make it work just like a control parameter.

Finally, good performance on tasks require the timing of the shifts in the at-

tractor landscape to be optimized, and that is exactly what is suggested by dynamic gating theory (O'Reilly and Frank 2006). As we reviewed in the motivation section, midbrain dopamine neurons fire precisely when the mind unexpectedly achieves motivational gains (either in externally delivered rewards or simply state improvements), and these moments are precisely when the prefrontal cortex would clamp down on the attractor landscape. Thus, the system would seem to flexibly shift r (and therefore the attractor landscape) on its own, according to changes in the motivational structure of the environment. For that reason, we could have titled this chapter "motivated executive control" – if dynamic gating theory is correct, then it is dopaminergic firing which governs the shifts in the attractor landscape defined over posterior cortex, motor cortex, subcortical regions, and the rest of the brain.

4.5 Closing thoughts

4.5.1 Distributed interactive control

Whereas dual systems theory holds that "controlling" is the property of a single privileged system (see for example Kahneman's table) we find that in an interactive system, multiple facets of control are distributed across multiple neurobiological structures. Recall Fuster's (2000) statement, "Pursuing methodological neatness, we have often been misled to the localization of cognitive functions that are

not localizable.” Similarly, here we see that attempts to localize control face immediate “who’s the boss?” problems. From a dynamical systems perspective, control is a fractionated rather than unitary concept, so these same issues with localization simply reproduce themselves during any attempt to “theoretically localize” control into a dual systems framework. We make two arguments here. First of all, remember our argument that “control must be responsive” (from the introduction, based on arguments of optimal control). However, prefrontal cortical representations are clamped for the duration of a task set (and similarly from the other perspectives on flexible top-down control – the dynamical systems control parameter is invariant to the evolution of state; the working memory processing requirement is for active maintenance). The restriction would severely impoverish the prefrontal cortex’s ability to control, because the clamping would force the prefrontal cortex into open-loop policies, which, as we argued in the introduction, are quite impotent. However, luckily the actor (in the basal ganglia), by virtue of being unclamped, remains free to be responsive to state, and in fact, according to a hierarchical reinforcement learning approach, the actor would be in charge of selecting new prefrontal representations whenever necessary based on the state. Thus, in a certain sense, the actor is in charge of the prefrontal cortex. (And in fact, pursuing this logic even further, the critic is in charge of the actor). To make a more complete argument about the distributed nature of control, consider once again the selective visual attention tasks related to the LIP. In these tasks as well, it could be asked *where* the control occurs. On the one hand, the LIP appears to implement the control, since it is bringing its neural firing patterns back to the

desired location in the space of locations for spatial attention. Yet, it could always be stipulated that some other region, higher up in the functional hierarchy, is controlling the LIP – perhaps for instance the PFC. From a dynamical systems perspective, we can decompose “flexible top-down control” into multiple aspects – first, there is the dynamical attraction (the very process which reestablishes the dominance of the top-down goal over pesky unintended distractors). This dynamical attraction process occurs in the LIP. Second, there is the prior selection of the attractor landscape. This landscape selection process occurs in the dorsolateral BG. Third, there is the adaptive critic which, through feedback, is responsible for teaching the actor how to best select the attractor landscapes. This critiquing process occurs in the ventral BG. Fourth, there is the actual presence of the reward structure in the environment, which governs this entire process. The primary rewards occur in lateral hypothalamus. Finally, there is the maintenance of the control parameter, which occurs in the prefrontal cortex. In other words, we see here that by investigating the dynamics, we have revealed a control by committee – each playing its specialized role in an interactive cooperative process – rather than control by dictatorial fiat, whereby some hot-shot intellectual superstar repeatedly bullying the classroom’s unrelenting dunce.

The distributed nature of control in an interactive dynamical system is illustrated in Figure 4.12

By investigating this control hierarchy, we are led from the decision layer by

way of prefrontal cortex through the subcortex, eventually back to the environment itself! According to the logical causal chain, the environment provides the basis for the structuring of the critique of the selections of the control parameter, and the control parameter determines the actual establishment of top-down control by transforming the irrelevant distractor. If a dual systems model wanted to cleave this network into a "control system" and an "automatic system," how would it do so? There is no obvious solution. The LIP provides the actual functional transformation from perturbation to goal-relevant action, and in that sense it would seem to be the executive controller. On the other hand, tracing the chain of command backwards five steps, we end up at the lateral hypothalamus! Hardly the typical proposal for a central executive. In short, the game doesn't make sense for a reason, because the control functionality requires these different pieces to fit into a unified puzzle in order for control to "work."

We cannot attribute psychological control in the brain to any single dominant head chef, because control simultaneously replicates itself at multiple spatial scales in a hierarchical interactive system. We start with one of our conclusions: whereas systems select states, control parameters select attractor landscapes. This notion recalls the fact that the functionality of control is commonly linked to "metacognition," as for example in Richard Petty's Metacognitive Model (Petty, Brinol, & DeMarree 2007). Because control parameters have their own dynamics, just along a superordinate time scale relative to the state variable, we could by induction make new dynamical systems out of the control parameters, whereby some meta-control parameter governs the dynamical selection of the control pa-

rameters. Given the hierarchical view of the PFC itself (Reynolds and O'Reilly 2009; Fuster 1997) there is reason to believe that these meta-dynamical systems are, in turn, nested within yet further higher-order dynamical systems. The pursuit would be a natural one, given the basic brain organizational principle of "loops within loops" (Merker 2004). Thus, while some researchers have called for a theory developing an understanding of the cooperations between multiple distinct dynamical networks (Ganguli et al. 2008), a very important instantiation of this endeavor would be to develop a theory of "nested dynamical systems," and this may mean that the brain sciences are requiring the pursuit of new mathematical frameworks in the field of applied dynamical systems. In fact, so long as a dynamical system has multiple control parameters, and therefore multiple nestings at potentially multiple levels, the possibility stands to developing a dynamical systems approach to linguistic functionalities like variable argument binding and recursive transformations. In this way, dynamical systems could satisfy Fodor and Pylyshyn (1988)'s influential complaint that connectionist networks lack the "systematicity" evident in human language – e.g., that it is impossible to understand "John loves Mary" without understanding "Mary loves John." Moreover, we can also pursue our argument in the other direction, towards the periphery, recalling the model from the "selection" chapter on how the brain keeps the eyes still (Seung 1996; Polk, Simen, Lewis & Freedman 2002).. In that model, the "controller" for the eyes are the premotor neurons medial vestibular nucleus (MVN) and the prepositus hypoglossi (PH). Although that system is believed to be a feedforward system, if we ask what controls the premotor neurons, we arrive at another dy-

namical system with another control parameter. Thus, while the “control” of eye gaze direction can be posed as a microcosm of our dynamical systems framework, that does not belittle the theory, but rather points towards a fundamental fact about “the fractal nature of control.”

4.5.2 The “dynamic projections” hypothesis of psychological control

A particularly fascinating variant of this theoretical framework is that the control parameter in the prefrontal cortex defines a projection of the higher-dimensional posterior cortical space. In other words, the prefrontal cortex causes a projection of the high-dimensional dynamics of posterior cortical space onto small regions of the space – i.e. “subspaces” or “manifolds” – that would be adaptive for the particular task at hand. (Note that this hypothesis is a special case of the “attractor landscape selection” hypothesis; here the prefrontal cortex would need to select a special kind of attractor landscape such that all the interesting competitive dynamics would occur in only small regions of posterior cortical space.) Thus, we call this hypothesis the “manifold selection” hypothesis, or alternatively, the “dynamic projections” hypothesis about psychological control.

Recent work in neuroscience has suggested that the prefrontal cortex may enact top-down control by implementing subspace projections in the posterior cortex. For example, although LIP-space contains high-dimensional popula-

tion codes for devoting attention to different spatial locations, neurophysiological recordings have revealed that controlled attention and integrative decision-making tasks cause LIP to exhibit low-dimensional dynamics (Ganguli et al. 2008). In particular, imagine that LIP neurons are firing stably at some attractor point in space. When a visual distractor appears, the neurons go haywire. But when that set of LIP neurons takes its state-space trajectory back towards the original attractor, it always travels back to the attractor along a single dimension. Dynamic movement along this particular dimension is meaningful, as it precisely reflects the amount of slowly integrating evidence for a perceptual decision. Although the paper did not investigate the dynamic projections hypothesis directly, it seems reasonable that, as the task goals change (and participants must maintain attention at different locations), the prefrontal cortex must be projecting the dynamics onto different single dimensions cutting through the high-dimensional space.

We find further support for the dynamic projections hypothesis by investigating once again the simulations of Rougier et al. (2005). The researchers compared the performance of a single interacting system models with multiple interacting system models on cognitive control tasks, such as the WCST and the Stroop task. They found that a multiple interacting systems model which included a basal ganglia component for learning motivational value, a prefrontal cortex component for maintaining the current goal representation, and dynamic gating between them outperformed the single interactive system models. Moreover, even though the additional components were designed simply to implement the specialized properties of the prefrontal cortex (active maintenance and hierarchical recurrent

connectivity) and basal ganglia (adaptive criticism and dynamic gating), these components interacted to construct a fascinating emergent property: rather than representing specific features of stimuli (e.g., blue), the prefrontal cortical neurons ended up representing content-less (i.e., more abstract) dimensions. For instance, after training on the Stroop and WCST tasks, the prefrontal neurons end up representing "shape", or "color", or "size", without specifying which shape or color or size. In this way, the prefrontal cortex ended up performing the syntactic binding of the rule-based symbolic computations – the assignment of variables to roles believed (Fodor & Pylyshyn 1988; Pinker 1997) to be the *sin qua non* of human language and symbolic logic thought. However, the prefrontal cortex performed these functions within a dynamic self-organizing system whose structure is parallel, distributed, and network-like. Thus, the multiple interacting systems models can implement "rule-based behavior" without losing interactivity or its non-symbolic, parallel, and distributed processing properties.

Now let us look more deeply at how the prefrontal cortex was affecting the posterior cortex in the Rougier et al. (2005) simulation. Their posterior cortex represented each stimulus inside a 145-unit layer. Thus, a medium-sized blue square stimulus with a given texture and location would have a 145-dimensional representation. (That is, the stimulus would be represented by a particular pattern of neural firing rates over 145 neuronal units, and thus would be captured by a point in a 145-dimensional Euclidean space). However, when the prefrontal cortex in that model was maintaining the strategic goal relevant to the current trial of the WCST (e.g. to categorize by color), it seemed to be effectively project-

ing the posterior cortical representation onto a smaller dimensional subspace (e.g. 19-dimensional representation). The smaller-dimensional representation would preserve the posterior cortical information relevant for the current goal – e.g. the stimulus’ color – while discarding information irrelevant to the current goal – e.g. the stimulus’ size, shape, texture, and location. But as the agent’s strategic goals flexibly changed over the course of the task, the prefrontal cortex flexibly shifted its projections. That is, the prefrontal cortex component, through dynamic gating from the basal ganglia, flexibly switched between different lower-dimensional representations (color or shape or location) of the same high-dimensional stimuli. In fact, these low-dimensional projections seem to be precisely what are responsible for the effective performance of the multiple systems model in Rougier et al (2005), who found strong correlations between (a) the model’s performance on strategic flexible goal tasks and (b) how well the prefrontal cortex component learned orthogonal dimensions that it could feedback into the posterior cortex.

Thus, we propose that the prefrontal cortex may be adaptively selecting a lower-dimensional subspace upon which to project the high-dimensional representations of the posterior cortex. The dynamics of mental processing (e.g. competition between different candidate mental representations or behavioral choices) would occur along these low dimensional subspaces, and thus the dynamic competition between mental representations would take different courses under different goals. While the mathematics of this proposal do not appear to be completely worked out (Steve Strogatz, personal communication), it is known that the reduction to a subspace does frequently results in a well-defined vector field

so long as the system contains an attracting invariant manifold (John Guckenheimer, personal communication).

The “dynamic projections” hypothesis provides an interesting, tractable theoretical explanation for social psychological phenomena involving the executive control of the implicit mind. For example, research in evaluative readiness has demonstrated that a deliberately adopted strategic goal – such as the desire to win a game or to perform well academically – can moderate implicit attitudes within a matter of milliseconds (Ferguson & Bargh 2004; Ferguson 2008) . An outstanding theoretical problem is: how can we explain evaluative readiness? That is, how do flexible strategic goals influence implicit attitudes? Evaluative readiness suggests a great deal of fluidity between flexible strategic goals and implicit attitudes. For instance, when an undergraduate student decides to head for the library to study for a final exam, her temporarily adopted strategic goal would cause her to experience greater implicit positivity towards libraries, books, and parties; however, just minutes later, if the student decides to step outside for a smoke break to make new friends, her newly adopted strategic goal would cause her to experience greater implicit positivity towards a cigarette. Current theoretical models in social psychology do not readily explain this “strategic fluidity” – the transient nature of strategic goals and their influence on implicit attitudes (for further discussion, see Ferguson and Wojnowicz 2011).

To help us approach this strategic fluidity, let us briefly revisit how implicit attitudes are measured in the evaluative readiness literature. A person’s evalu-

ation of "cigarettes" for example is typically measured based on reaction times to positive words such as "good" after being primed with a stimulus word like "cigarette." From the perspective of most connectionist networks with distributed representations, reaction times are a proxy for proximity in state space (how long it takes the mind to transition between distributed patterns) (e.g., Cree, McRae, & McNorgan 1999). In other words, more similar distributed patterns would show stronger priming. Thus, when a person wants a cigarette, the distributed pattern for "cigarette" and the distributed pattern for positive concepts like "good" should look more similar than when a person doesn't want to cigarette. This means that a person's currently operating strategic goal should modulate the representation of concepts! How might this be happening?

We argue here that the "dynamic projections" hypothesis provides an excellent mechanism for understanding evaluative readiness. That is, we hold that the prefrontal cortex seems to be adaptively selecting a lower-dimensional subspace upon which to project the high-dimensional representations of the posterior cortex. The selection of the subspace depends upon the currently operating strategic goal (studying vs. socializing). When the prefrontal cortex chooses a different strategic goal, it actively maintains a different firing pattern, which thereby projects the posterior cortical representations onto a different subspace. This projection mechanism would explain evaluative readiness, because priming times (reflecting evaluations) would depend upon the projection onto various lower-dimensional subspaces.

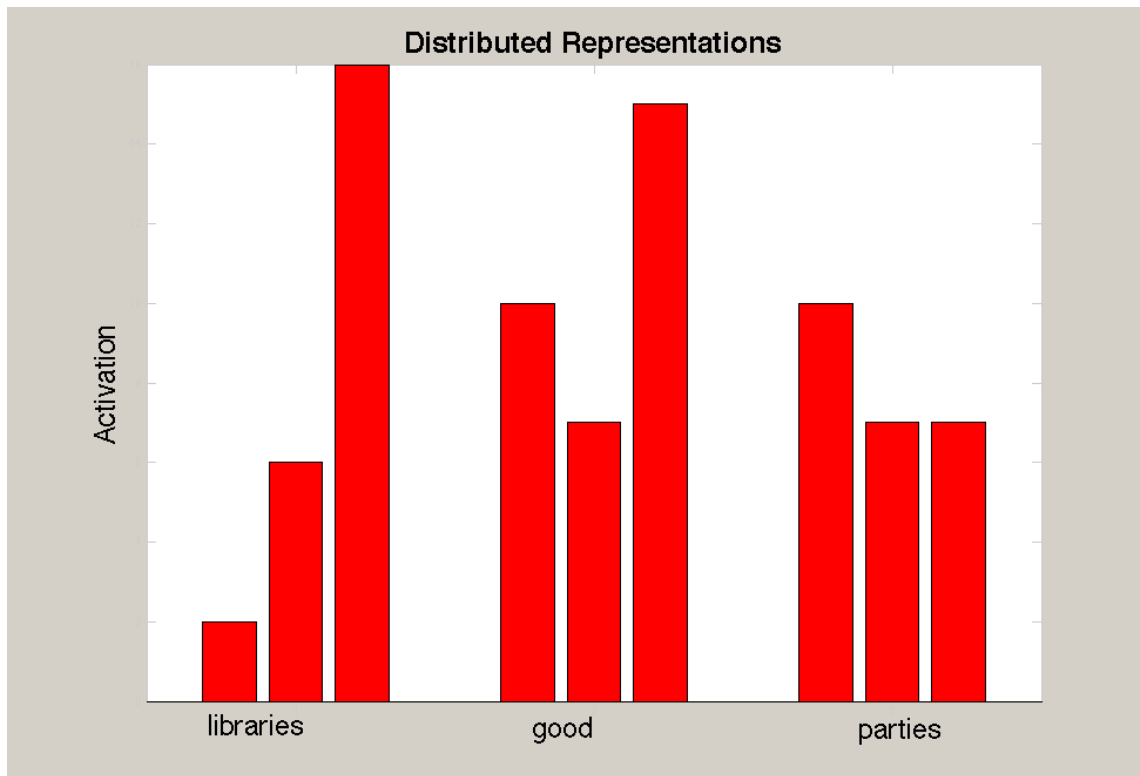


Figure 4.13: Hypothetical distributed representations for libraries, good, and parties

For a simple cartoon example, imagine a posterior cortex with only three neurons. Then, the "high-dimensional" posterior cortical representations for semantic concepts would have 3 dimensions, and lower-dimensional projections would have fewer dimensions (perhaps 2). Figure 4.13 shows some hypothetical distributed representations for libraries, good, and parties in the hypothetical three-neuron posterior cortex. Figure 4.14 depicts these same three hypothetical representations in three-dimensional Euclidean space, where $(x,y,z) = (-4,0,10)$ for "par-

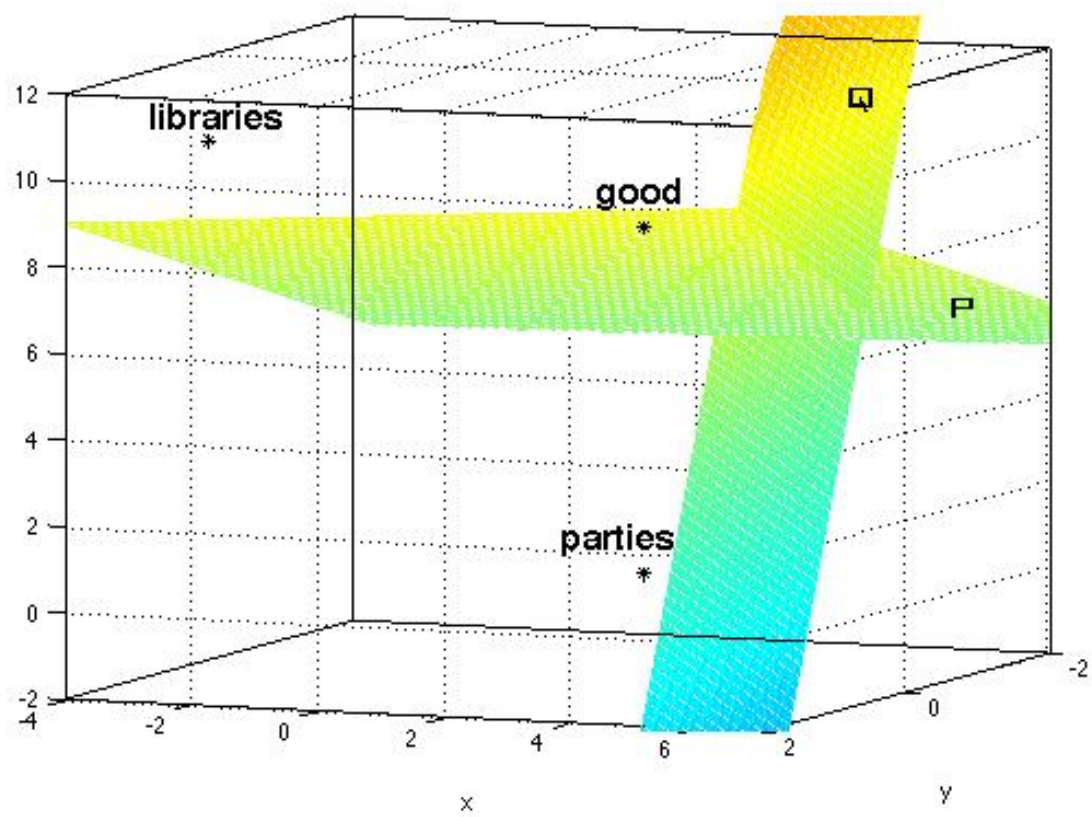


Figure 4.14: The relative positivity of libraries vs. parties depends upon how the prefrontal cortex projects high-dimensional representations onto low-dimensional subspaces

ties", (4, 1,9) for "libraries", and (4,1,1) for "good." Whether the mental representations for "libraries" and "good things" are close to each other or far depends on how the prefrontal cortex projects the pattern-formation dynamics onto subspaces. If the projection occurs along the plane Q, then "libraries" primes "good" very strongly, much more strongly than "parties." If the projection occurs along the plane P, "libraries" would not prime "good" very strongly, especially in comparison to "parties."

The main point is that some recent multiple interacting systems models suggest that – based on the distinguishing properties of the prefrontal cortex and basal ganglia – dimensionality projections are what allows the prefrontal cortex to achieve flexible cognitive control in accordance with current motivations. In the case of the Rougier et al. (2005) model of the WCST, the prefrontal cortex provided a mechanism through which the model could flexibly shift around the similarity groupings of blue squares vs. green squares vs. green circles, in accordance with task demands. We can construe evaluative readiness in exactly the same way – the prefrontal cortex provides a mechanism through which the posterior cortex flexibly shifts around evaluations of books and libraries vs. parties and beers depending on current goals.

CHAPTER 5

CONCLUSION

5.1 Summary of Dissertation

In summary, this dissertation attempted to support a dynamical interactive approach to mental processing, particularly in social contexts.

The empirical chapters (Chapters 2 and 3) investigated the dynamics of mental processing by recording hand-movement trajectories during an evaluative decision-making task. Despite the fact that these evaluative decisions were made explicitly, in both chapters, we found evidence that the decision-making dynamics were biased by informational sources that were implicit. In Chapter 2, the sources of informational bias were negative racial associations, which were assumed to be aversive and implicit based on previous research. In Chapter 3, the sources of informational bias were subliminal conditioning, which can be regarded more directly and self-evidently as implicit. These findings support theories of mental processing that involve a single set of interactive computational principles, and would not be as elegantly predicted or explained by dualistic theories of mental processing, which have not yet (to date) provided a dynamical account of the explicit mind, or of how the explicit mind would interact with the implicit mind.

The theoretical chapter (Chapter 4) attempted to push forward by developing a theory of psychological control that arises out of a single set of interactive

computational principles. In that chapter, we investigated three frequently mentioned capacities that a psychological controller should have (decisive selection, strategic goal pursuit, and top-down processing flexibility), and we determined how an interactive system could accomplish these functions. We described these control capacities with respect to a dynamical systems framework, so that the dualistic notion of a separated homunculus making a decisive selection, striving for distant end states, or shifting goals on the spot could be replaced by tangible alternative concepts representing how these phenomena could work in a dynamical neural system.

5.2 Limitations

A major limitation of the empirical studies is that they do not themselves fully address a more recent resurgent debate in the literature about whether hand-movement trajectories in competitive decision-making tasks reflect a continuous evolution of the decision-making process (as argued by Spivey, Knoblich, Dale, & Grosjean 2010) or the results of a discrete decision-making process, whereby a succession of two discrete motor commands are superimposed onto each other (as argued by van der Wel, Eder, Mitchel, Walsh, & Rosenbaum 2009). These research groups have both constructed computer simulations that generate continuously curved hand-movement trajectories from the polynomial equations of Henis and Flash (1995) (which describe how hand-movement trajectories move

over time from an initial (X,Y) coordinate to a desired (X,Y) coordinate). However, the models make very different assumptions about processing mechanisms. Spivey et al (2010) construct hand-movement trajectories based on the assumption of a continuously evolving decision-making process. They construct hand-movements through the superposition of two commands – a straight-line to the response and a straight line to the competitor – whose weights depend upon the continuously evolving activation levels of the underlying interpretation at any given time. In contrast, van der Wel et al (2010) construct hand-movement trajectories by assuming an initial completely vertical motion (planned to move from the start box to midpoint between the targets) followed by an initial completely horizontal motion (planned to move from the midpoint between the targets to the chosen target). In the van der Wel et al (2010) model, the temporal evolutions of the successive discrete movements are also governed by the Henis and Flash (1995) equations, but the two movements are superimposed when they overlap, thus creating trajectories which look continuously curved. Because the second movement is presumed to begin later on average in the cohort condition than in the control condition, the cohort condition's trajectories curve out more towards the competing response, even though the underlying mechanism is discrete (i.e., merely a delayed discrimination).

Spivey et al (2010) has made the case that previous mouse-tracking data has supported continuous decision-making process on the basis of two empirical arguments that I will review here. First, hand movement trajectories curve towards competitor response options even when the chosen response is located completely

vertically upward. Second, hand movement trajectories in many data sets (including some unpublished data from the subliminal dynamics project described in Chapter 3) show early deflections (in angles of departure from the stimulus box), which the discrete model cannot generate. While the present writer agrees that these observations favor the continuous processing model over the discrete processing model, it is true that the data incorporated in the current dissertation do not directly address these concerns, which is a clear limitation. In the ideal case, since this dissertation's empirical studies cannot resolve these theoretical questions in and of themselves, further research would disentangle whether the observed trajectories towards "like" and "dislike" evaluations are generated by a continuous or discrete decision-making process.

Another limitation of the empirical studies is the experimental logic in Chapter 2 motivating the design and interpretation of Study 2. Subsequent to publication of that research, reviewers of this dissertation have pointed out the fallacy in our usage of the lack of significance in one context to make an argument about the difference in contexts. That is, in Chapter 2, we argued in the following way: conditions A and B are significantly different in experiment 1 and not significant in experiment 2, "Thus the results of experiment 1...." In actuality, one must demonstrate a significant interaction for such an argument to be made. Thus, subsequent to publication of the research, we have attempted to explore support for such an interaction. We ran a linear regression of maximum hand movement deviation on study and stimulus (black people vs. white people), including an interaction term for study and stimulus. We used robust standard errors and clustering on

subject id to account for possible heteroscedasticity between subgroups and for correlation between measurements on the same subject. Comparing Study 2 and Study 1, we found that the interaction term study*stimulus approached marginal significance, $t(121)=1.59$, $p=.11$. Comparing Study 3 and Study 1, we found that the interaction term study*stimulus was significant, $t(123)=3.18$, $p=.002$. However, the latter statistical investigation is inconclusive; for instance, the significant interaction term in comparing Study 3 and Study 1 could have been driven by the difference in stimulus labeling (e.g. African Americans vs. black people) rather than the difference in competitor response (chemical vs. dislike). Future research should provide a more appropriate and conclusive investigation of the role of the competing response box in eliciting hand-movement curvature.

At the theoretical level, the arguments of the dissertation have not strayed very far from the very elementary level of a qualitative debate about whether mental processing should be construed as a dynamical system or not. In this dissertation, the concept of dynamical systems used has been the mathematical concept, whereby a dynamical system is a set of coupled differential equations (see for example Hirsch, Smale, & Devaney 2004's textbook *Differential equations, dynamical systems, and an introduction to chaos*; or Guckenheimer & Holmes 1990's textbook *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*). As has been pointed out, the claim "system x should be viewed as a dynamical system" is an extremely weak claim, as nearly any observable process in the world could be described through a dynamical systems model. However, there are merits of defending even this weak claim. In particular, there are some necessary features of a

non-trivial dynamical systems model; these would include (a) interactivity (i.e. at least some of the components of the system interact) and (b) numerical representation (i.e the state of the components can be meaningfully represented by a scalar or vector). Theoretical approaches to cognition in social psychology (especially dual systems models) are frequently NOT dynamical systems. Some dual systems approaches fail necessary condition (a) (see e.g., Rydell & McConnell 2006); they posit rule-based and associative systems which do not interact. Other dual systems approaches fail necessary condition (b) (see e.g. Gawronski & Bodenhausen 2006); they posit rule-based and associative systems which are "allowed" to interact, but they do not have components whose states have numerical representations. Thus, although the claim "mental processing should be understood as a dynamical system" is very weak, it meaningfully excludes large classes of theoretical models in social psychological psychology. Thus, we believe that the primary thesis of this dissertation – that mental processing should be considered as a dynamical system – is important and meaningful. The dissertation favors building social psychology theory out of biologically plausible models of cognition which are largely popular outside social psychology (such as O'Reilly's LEABRA models; or Spivey's Normalized Recurrence model). Since these models use population codes to explain apparently rule-based phenomena, they satisfy necessary condition (b), and because of the principles of mutual inhibition, reciprocal feedback, and partial cascading, they satisfy necessary condition (a). However, since the dissertation has devoted much of its effort towards defending the very preliminary argument that mental processing should be considered as a dynamical

system, rather than not, there remains an incredible amount of follow-up work to explore and test the many implications for typical social psychological experiments and theory.

Along those lines, a major limitation of the theoretical chapter is that the theoretical ideas need to be channeled into clear empirical predictions for psychological control tasks in social psychological (or other) settings. I hope that future research may be aided by this dissertation's attempt to describe how psychological control phenomena could arise through dynamic interactions; that is, that the mind flexibly selects motivated attractor landscapes. Providing empirical corroboration of some of the main ideas would greatly bolster the argument. In particular, it would be interesting to devise an empirical study that would seek support for the most strongly stated claim – that of the dynamic projections hypothesis. The dynamic projections hypothesis subsumes the other three "theses;" thus empirical research supporting the dynamic projections hypothesis would also provide solid support for the chapter's broad argument about psychological control from an interactive perspective. However, at the current moment in time, collectable behavioral data (such as mouse-tracking, reaction-time tests, etc) relate to these more fine-grained theoretical proposals of the dynamical systems approach in frustratingly sketchy and tenuous ways. It is my hope that future research will continue to either develop more fine-grained measurement techniques, or to devise more clever ways to use the current measurement techniques.

BIBLIOGRAPHY

- [1] D. Albarracin, B.T. Johnson, and M.P. Zanna. *The handbook of attitudes*. Lawrence Erlbaum, 2005.
- [2] BJ Baars. How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? In *CIBA Foundation Symposium*, volume 174, page 282, 1993.
- [3] T.E. Baker and C.B. Holroyd. Which Way Do I Go? Neural Activation in Response to Feedback and Spatial Processing in a Virtual T-Maze. *Cerebral Cortex*, 19(8):1708, 2009.
- [4] E. Balceris and D. Dunning. See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, 91(4):612–625, 2006.
- [5] I. Bar-Gad, G. Morris, and H. Bergman. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71(6):439–473, 2003.
- [6] J.A. Bargh. The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. *Handbook of social cognition: Basic processes*, 1:1–40, 1994.
- [7] J.A. Bargh and M.J. Ferguson. Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin*, 126(6):925–945, 2000.
- [8] J.A. Bargh, P.M. Gollwitzer, A. Lee-Chai, K. Barndollar, and R. Trötschel. The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, 81(6):1014–1027, 2001.
- [9] A.G. Barto. "1" 1 adaptive critics and the basal ganglia. *Models of information processing in the basal ganglia*, page 215, 1994.

- [10] R.F. Baumeister, E. Bratslavsky, C. Finkenauer, and K.D. Vohs. Bad is stronger than good. *Review of general psychology*, 5(4):323–370, 2001.
- [11] R.F. Baumeister, K.D. Vohs, and D.M. Tice. The strength model of self-control. *Current Directions in Psychological Science*, 16(6):351, 2007.
- [12] W. Bechtel and A. Abrahamsen. *Connectionism and the Mind*. Blackwell, 1991.
- [13] R. Bellman. The theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716–719, 1952.
- [14] R. Bellman and S. Dreyfus. *Applied dynamic programming*. RAND, 1962.
- [15] R.J. Beninger, F. Bellisle, and P.M. Milner. Schedule control of behavior reinforced by electrical stimulation of the brain. *Science*, 196(4289):547, 1977.
- [16] K. Berridge and P. Winkielman. What is an unconscious emotion?(the case for unconscious” liking”). *Cognition & Emotion*, 17(2):181–211, 2003.
- [17] K.C. Berridge and T.E. Robinson. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3):309–369, 1998.
- [18] C.J. Berry, D.R. Shanks, and R.N.A. Henson. A single-system account of the relationship between priming, recognition, and fluency. *Journal of Experimental Psychology-Learning Memory and Cognition*, 34(1):97–110, 2008.
- [19] D.P. Bertsekas. *Dynamic programming and stochastic control*, volume 125. Academic Pr, 1976.
- [20] D.P. Bertsekas et al. *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995.
- [21] G. Bi and M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience*, 18(24):10464–10472, 1998.

- [22] J.W. Bisley and M.E. Goldberg. Neuronal activity in the lateral intraparietal area and spatial attention. *Science*, 299(5603):81, 2003.
- [23] J.W. Bisley and M.E. Goldberg. Neural correlates of attention and distractibility in the lateral intraparietal area. *Journal of neurophysiology*, 95(3):1696, 2006.
- [24] G.V. Bodenhausen. Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science*, 1(5):319, 1990.
- [25] R. Bogacz and K. Gurney. The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural computation*, 19(2):442–477, 2007.
- [26] M.M. Botvinick, Y. Niv, and A.C. Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.
- [27] P. Briñol, RE Petty, and MJ McCaslin. Changing attitudes on implicit versus explicit measures: What is the difference. *Attitudes: Insights from the new implicit measures*, pages 285–326, 2009.
- [28] J.C. Brunstein, O.C. Schultheiss, and R. Grassmann. Personal Goals and Emotional Well-Being: The Moderating Role of Motive Dispositions* 1. *Journal of Personality and Social Psychology*, 75(2):494–508, 1998.
- [29] J.R. Busemeyer and J.T. Townsend. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100:432–432, 1993.
- [30] L. Chelazzi, E.K. Miller, J. Duncan, and R. Desimone. A neural basis for visual search in inferior temporal cortex. *Nature*, 363(6427):345–347, 1993.
- [31] N. Chomsky. A review of BF Skinner’s Verbal Behavior. *Readings in language and mind*, pages 413–441, 1996.
- [32] N. Chomsky. *Language and mind*. Cambridge Univ Pr, 2006.

- [33] N. Chomsky and G.A. Miller. Finite state languages*. *Information and control*, 1(2):91–112, 1958.
- [34] M.H. Christiansen and N. Chater. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205, 1999.
- [35] P. Cisek and J.F. Kalaska. Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron*, 45(5):801–814, 2005.
- [36] A. Cleeremans. Attention and awareness in sequence learning. In *Proceedings of the fifteenth annual conference of the Cognitive Science Society: June 18 to 21, 1993, Institute of Cognitive Science, University of Colorado, Boulder*, page 330. Lawrence Erlbaum, 1993.
- [37] T.A. Cleland, B.A. Johnson, M. Leon, and C. Linster. Relational representation in the olfactory system. *Proceedings of the National Academy of Sciences*, 104(6):1953, 2007.
- [38] A.M. Collins and E.F. Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407–428, 1975.
- [39] F.R. Conrey and E.R. Smith. Attitude representation: Attitudes as patterns in a distributed, connectionist representational system. *Social Cognition*, 25(5):718–735, 2007.
- [40] G.S. Cree, K. McRae, and C. McNorgan. An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414, 1999.
- [41] W.A. Cunningham, M.K. Johnson, C.L. Raye, J.C. Gatenby, J.C. Gore, and M.R. Banaji. Separable neural components in the processing of black and white faces. *Psychological Science*, 15(12):806, 2004.
- [42] R. Dale, C. Kehoe, and M.J. Spivey. Graded motor responses in the time course of categorizing atypical exemplars. *Memory & cognition*, 35(1):15, 2007.

- [43] M. Davis. The role of the amygdala in fear and anxiety. *Annual review of neuroscience*, 15(1):353–375, 1992.
- [44] N.D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005.
- [45] J.W. De Fockert, G. Rees, C.D. Frith, and N. Lavie. The role of working memory in visual selective attention. *Science*, 291(5509):1803, 2001.
- [46] L.T. DeCarlo. On the meaning and use of kurtosis. *Psychological Methods*, 2(3):292–307, 1997.
- [47] D.C. Dennett and D.C. Dennett. *Brainstorms: Philosophical essays on mind and psychology*. The MIT Press, 1981.
- [48] R. Desimone. Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1245–1255, 1998.
- [49] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.
- [50] P.G. Devine. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5–18, 1989.
- [51] J.F. Dovidio, K. Kawakami, and S.L. Gaertner. Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82(1):62–68, 2002.
- [52] D. Durstewitz, M. Kelc, and O. Gunturkun. A neurocomputational theory of the dopaminergic modulation of working memory functions. *Journal of Neuroscience*, 19(7):2807, 1999.
- [53] J.L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

- [54] J.L. Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2):195–225, 1991.
- [55] W. Erlhagen and G. Schoner. Dynamic field theory of movement preparation. *Psychological Review*, 109(3):545–572, 2002.
- [56] T.A. Farmer, S.E. Anderson, and M.J. Spivey. Gradiency and visual context in syntactic garden-paths. *Journal of memory and language*, 57(4):570–595, 2007.
- [57] R.H. Fazio. Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in experimental social psychology*, 23(75-109), 1990.
- [58] R.H. Fazio, J.R. Jackson, B.C. Dunton, and C.J. Williams. Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline?. *Journal of Personality and Social Psychology*, 69(6):1013–1027, 1995.
- [59] R.H. Fazio and M.A. Olson. Implicit Measures in Social Cognition Research: Their Meaning and Use. *Annual review of psychology*, pages 297–328, 2003.
- [60] S. Fein and S.J. Spencer. Prejudice as self-image maintenance: Affirming the self through derogating others. *Journal of Personality and Social Psychology*, 73(1):31, 1997.
- [61] M.J. Ferguson and J.A. Bargh. Liking Is for Doing: The Effects of Goal Pursuit on Automatic Evaluation. *Journal of Personality and Social Psychology*, 87(5):557–572, 2004.
- [62] M.J. Ferguson and J.A. Bargh. Beyond the attitude object: Automatic attitudes spring from object-centered-contexts. *Implicit measures of attitudes*, pages 216–246, 2007.
- [63] M.J. Ferguson and J.A. Bargh. Immediacy and Automaticity of Evaluation Evaluative Readiness: The Motivational Nature of Automatic Evaluation. *Handbook of approach and avoidance motivation*, page 289, 2008.

- [64] M.J. Ferguson and M.T. Wojnowicz. Evaluative Readiness: Exploring the how's and why's. *Social and Personality Psychology Compass*, accepted.
- [65] S.T. Fiske. Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, 38(6):889–906, 1980.
- [66] J.A. Fodor and Z.W. Pylyshyn. Connectionism and cognitive architecture: a critical analysis, Connections and symbols. *A Cognition Special Issue*, S. Pinker and J. Mehler (eds.), MIT Press, Cambridge, MA, pages 3–71, 1988.
- [67] S. French. *Sequencing and scheduling: an introduction to the mathematics of the job-shop*. Halsted Press, 1982.
- [68] J.M. Fuster. Network memory. *Trends in Neurosciences*, 20(10):451–459, 1997.
- [69] J.M. Fuster. Executive frontal functions. *Experimental Brain Research*, 133(1):66–70, 2000.
- [70] S. Galdi, L. Arcuri, and B. Gawronski. Automatic mental associations predict future choices of undecided decision-makers. *Science*, 321(5892):1100, 2008.
- [71] S. Ganguli, J.W. Bisley, J.D. Roitman, M.N. Shadlen, M.E. Goldberg, and K.D. Miller. One-dimensional dynamics of attention and decision making in LIP. *Neuron*, 58(1):15–25, 2008.
- [72] B. Gawronski and G.V. Bodenhausen. Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5):692, 2006.
- [73] D.T. Gilbert, E.C. Pinel, T.D. Wilson, S.J. Blumberg, and T.P. Wheatley. Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75(3):617, 1998.
- [74] P. Glasserman. *Monte Carlo methods in financial engineering*. Springer Verlag, 2004.

- [75] J.I. Gold and M.N. Shadlen. Representation of a perceptual decision in developing oculomotor commands. *Nature*, 404(6776):390–394, 2000.
- [76] J.I. Gold and M.N. Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30:535, 2007.
- [77] J.P. Gottlieb, M. Kusunoki, and M.E. Goldberg. The representation of visual salience in monkey parietal cortex. *Nature*, 391(6666):481–484, 1998.
- [78] A.R. Green, D.R. Carney, D.J. Pallin, L.H. Ngo, K.L. Raymond, L.I. Iezzoni, and M.R. Banaji. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine*, 22(9):1231–1238, 2007.
- [79] A.G. Greenwald, D.E. McGhee, J.L.K. Schwartz, et al. Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, 74:1464–1480, 1998.
- [80] A.G. Greenwald and B.A. Nosek. Attitudinal dissociation: What does it mean. *Attitudes: Insights from the new implicit measures*, pages 65–82, 2008.
- [81] D.L. Hamilton and R.K. Gifford. Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, 12(4):392–407, 1976.
- [82] TE Hazy, MJ Frank, and RC O'Reilly. Banishing the homunculus: making working memory work. *Neuroscience*, 139(1):105–118, 2006.
- [83] F. Heider. Attitudes and cognitive organization. *Journal of Psychology*, 21(1):107–112, 1946.
- [84] E.A. Henis and T. Flash. Mechanisms underlying the generation of averaged modified trajectories. *Biological Cybernetics*, 72(5):407–419, 1995.
- [85] E.T. Higgins. Knowledge activation: Accessibility, applicability, and salience. *Social psychology: Handbook of basic principles*, pages 133–168, 1996.

- [86] W. Hofmann, W. Rauch, and B. Gawronski. And deplete us not into temptation: Automatic attitudes, dietary restraint, and self-regulatory resources as determinants of eating behavior. *Journal of Experimental Social Psychology*, 43(3):497–504, 2007.
- [87] J.C. Houk, J.L. Adams, and A.G. Barto. A model of how the basal ganglia generate and use neural signals that predict reinforcement. *Models of information processing in the basal ganglia*, pages 249–270, 1995.
- [88] A.J. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. *Advances in neural information processing systems*, pages 1547–1554, 2003.
- [89] E.T. Jaynes and G.L. Bretthorst. *Probability theory: the logic of science*. Cambridge Univ Pr, 2003.
- [90] D. Joel, Y. Niv, and E. Ruppín. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, 15(4-6):535–547, 2002.
- [91] C.H. Jordan, S.J. Spencer, and M.P. Zanna. Types of high self-esteem and prejudice: How implicit self-esteem relates to ethnic discrimination among high explicit self-esteem individuals. *Personality and Social Psychology Bulletin*, 31(5):693, 2005.
- [92] C.M. Judd, R.A. Drake, J.W. Downing, and J.A. Krosnick. Some dynamic properties of attitude structures: Context-induced response facilitation and polarization. *Journal of Personality and Social Psychology*, 60(2):193–202, 1991.
- [93] D. Kahneman. Maps of bounded rationality: A perspective on intuitive judgment and choice. *Nobel Prize Lecture*, 8, 2002.
- [94] D. Kahneman and S. Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, pages 49–81, 2002.
- [95] S.A. Kauffman. *The origins of order*. Oxford University Press New York, 1993.

- [96] J.A.S. Kelso. *Dynamic patterns: The self-organization of brain and behavior*. The MIT Press, 1995.
- [97] E. Koechlin and C. Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235, 2007.
- [98] Z. Kunda and S.J. Spencer. When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, 129(4):522–544, 2003.
- [99] H.J. Kushner and P. Dupuis. *Numerical methods for stochastic control problems in continuous time*. Springer Verlag, 2001.
- [100] M.W. L Chee, N. Sriram, C.S. Soon, and K.M. Lee. Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. *NeuroReport*, 11(1):135, 2000.
- [101] V.A.F. Lamme and P.R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, 2000.
- [102] C.C. Lapsch, D. Durstewitz, L.J. Chandler, and J.K. Seamans. Successful choice behavior is associated with distinct and coherent network states in anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, 105(33):11963, 2008.
- [103] E.B. Lee and L. Markus. *Foundations of optimal control theory*. Wiley New York, 1967.
- [104] M.D. Lieberman. Social cognitive neuroscience: a review of core processes. *Annual Review of Neuroscience*, 58, 2006.
- [105] D. Ludwig, D.D. Jones, and CS Holling. Qualitative analysis of insect outbreak systems: the spruce budworm and forest. *The Journal of Animal Ecology*, 47(1):315–332, 1978.

- [106] R. Major and N. White. Memory facilitation by self-stimulation reinforcement mediated by the nigro-neostriatal bundle. *Physiology & Behavior*, 20(6):723–733, 1978.
- [107] J.R. Manns, M.W. Howard, and H. Eichenbaum. Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, 56(3):530–540, 2007.
- [108] O. Mazor and G. Laurent. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron*, 48(4):661–673, 2005.
- [109] J.L. McClelland. On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, 86(4):287–330, 1979.
- [110] J.L. McClelland, B.L. McNaughton, and R.C. O’Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–457, 1995.
- [111] M. McCloskey and N.J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *The psychology of learning and motivation: Advances in research and theory*, 24:109–165, 1989.
- [112] A.R. McConnell and J.M. Leibold. Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes* 1. *Journal of Experimental Social Psychology*, 37(5):435–442, 2001.
- [113] C. McKinstry, R. Dale, and M.J. Spivey. Action dynamics reveal parallel competition in decision making. *Psychological Science*, 19(1):22, 2008.
- [114] B. McMurray, M.K. Tanenhaus, R.N. Aslin, and M.J. Spivey. Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *Journal of Psycholinguistic Research*, 32(1):77–97, 2003.
- [115] K. McRae, M.J. Spivey-Knowlton, and M.K. Tanenhaus. Modeling the influ-

ence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38:283–312, 1998.

- [116] B. Merker. Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex*, 40:559–576, 2004.
- [117] B. Merker. Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30(01):63–81, 2007.
- [118] R.R. Metzger, N.T. Greene, K.K. Porter, and J.M. Groh. Effects of reward and behavioral context on neural activity in the primate inferior colliculus. *Journal of Neuroscience*, 26(28):7468, 2006.
- [119] E.K. Miller. The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, 1(1):59–65, 2000.
- [120] E.K. Miller and J.D. Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.
- [121] E.K. Miller and J.D. Cohen. An integrative theory of prefrontal cortex function. *Annual Reviews of Neuroscience*, 24:167–202, 2003.
- [122] E.K. Miller, C.A. Erickson, and R. Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, 16(16):5154, 1996.
- [123] E.K. Miller, L. Li, and R. Desimone. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience*, 13(4):1460, 1993.
- [124] J. Mirenowicz and W. Schultz. Importance of unpredictability for reward responses in primate dopamine neurons. *Journal of Neurophysiology*, 72(2):1024, 1994.
- [125] J. Mirenowicz and W. Schultz. Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli. *Nature*, pages 449–451, 1996.

- [126] B.R.C. Molesworth and B. Chang. Predicting pilots risk-taking behavior through an implicit association test. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 51(6):845–857, 2009.
- [127] J.L. Monahan, S.T. Murphy, and R.B. Zajonc. Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science*, 11(6):462, 2000.
- [128] S. Monsell and J. Driver. *Control of cognitive processes: Attention and performance XVIII*. The MIT Press, 2000.
- [129] P.R. Montague, P. Dayan, and T.J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of Neuroscience*, 16(5):1936–1947, 1996.
- [130] P.R. Montague, S.E. Hyman, and J.D. Cohen. Computational roles for dopamine in behavioural control. *Nature*, 431(7010):760–767, 2004.
- [131] H. Nakahara, S. Amari, and O. Hikosaka. Self-organization in the basal ganglia with modulation of reinforcement signals. *Neural computation*, 14(4):819–844, 2002.
- [132] R.E. Nisbett and T. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–258, 1998.
- [133] M.K. Nock and M.R. Banaji. Prediction of suicide ideation and attempts among adolescents using a brief performance-based test. *Journal of Consulting and Clinical Psychology*, 75(5):707, 2007.
- [134] B.A. Nosek. Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, 134:565–584, 2005.
- [135] K.S. O’Brien, J.D. Latner, J. Halberstadt, J.A. Hunter, J. Anderson, and P. Caputi. Do antifat attitudes predict antifat behaviors? *Obesity*, 16:S87–S92, 2008.
- [136] M.A. Olson and R.H. Fazio. Implicit attitude formation through classical conditioning. *Psychological Science*, 12(5):413–417, 2001.

- [137] R.C. O'Reilly. Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2(11):455–462, 1998.
- [138] R.C. O'REILLY, T.S. Braver, and J.D. Cohen. 11 a biologically based computational model of working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, page 375, 1999.
- [139] R.C. O'Reilly and M.J. Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2006.
- [140] R.C. O'Reilly, T.S. Braver, and J.D. Cohen. A biologically based computational model of working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, pages 375–411, 1999.
- [141] D.N. Pandya and C.L. Barnes. Architecture and connections of the frontal lobe. *The frontal lobes revisited*, pages 41–72, 1987.
- [142] O. Penrose. Foundations of statistical mechanics. *Reports on Progress in Physics*, 42:1937, 1979.
- [143] R.E. Petty. *Attitudes: Insights from the new implicit measures*. Psychology Pr, 2008.
- [144] R.E. Petty, P. Briñol, and K.G. DeMarree. The Meta-Cognitive Model (MCM) of attitudes: Implications for attitude measurement, change, and strength. *Social Cognition*, 25(5):657–686, 2007.
- [145] E.A. Phelps, K.J. O'Connor, W.A. Cunningham, E.S. Funayama, J.C. Gatenby, J.C. Gore, and M.R. Banaji. Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, 12(5):738, 2000.
- [146] S. Pinker. *How the mind works*. 1997. NY: Norton, 1997.
- [147] S. Pinker. *How the mind works*. WW Norton & Company, 1999.

- [148] T.A. Polk, P. Simen, R.L. Lewis, and E. Freedman. A computational approach to control in complex cognition. *Cognitive Brain Research*, 15(1):71–83, 2002.
- [149] R.F. Port and T. Van Gelder. *Mind as motion: Explorations in the dynamics of cognition*. The MIT Press, 1995.
- [150] J.N.J. Reynolds, B.I. Hyland, and J.R. Wickens. A cellular mechanism of reward-related learning. *Nature*, 413(6851):67–70, 2001.
- [151] J.N.J. Reynolds and J.R. Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4-6):507–521, 2002.
- [152] J.R. Reynolds and R.C. O'Reilly. Developing PFC representations using reinforcement learning. *Cognition*, 113(3):281–292, 2009.
- [153] J.A. Richeson and J.N. Shelton. When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science*, 14(3):287–290, 2003.
- [154] R.M. Roe, J.R. Busemeyer, and J.T. Townsend. Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2):370–392, 2001.
- [155] T.T. Rogers and J.L. McClelland. *Semantic cognition: A parallel distributed processing approach*. The MIT Press, 2004.
- [156] E.T. Rolls and M.J. Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, 73(2):713, 1995.
- [157] K.W. Ross. Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Operations Research*, 37(3):474–477, 1989.
- [158] N.P. Rougier, D.C. Noelle, T.S. Braver, J.D. Cohen, and R.C. O'Reilly. Prefrontal cortex and flexible cognitive control: Rules without symbols. *Pro-*

ceedings of the National Academy of Sciences of the United States of America, 102(20):7338, 2005.

- [159] P. Rozin and E.B. Royzman. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4):296, 2001.
- [160] R.J. Rydell, A.R. McConnell, D.M. Mackie, and L.M. Strain. Of Two Minds. *Psychological Science*, 17(11):954, 2006.
- [161] G. Schoner, P.G. Zanone, and JA Kelso. Learning as change or coordination dynamics: Theory and experiment. *Journal of Motor Behavior*, 24(1):29–48, 1992.
- [162] W. Schultz. Dopamine neurons and their role in reward mechanisms. *Current Opinion in Neurobiology*, 7(2):191–197, 1997.
- [163] W. Schultz. Getting formal with dopamine and reward. *Neuron*, 36(2):241–263, 2002.
- [164] N. Schwartz. Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in experimental social psychology*, 26:123–123, 1994.
- [165] N. Schwarz and G. Bohner. The construction of attitudes. *Blackwell handbook of social psychology: Intraindividual processes*, 1:436–457, 2001.
- [166] WB Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiat*, 20:11–21, 1957.
- [167] JA Sethian and A. Vladimirovsky. Ordered upwind methods for static Hamilton–Jacobi equations. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11069, 2001.
- [168] H.S. Seung. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339, 1996.

- [169] B. Seymour, J.P. O'Doherty, P. Dayan, M. Koltzenburg, A.K. Jones, R.J. Dolan, K.J. Friston, and R.S. Frackowiak. Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–667, 2004.
- [170] D.F. Sherry and D.L. Schacter. The evolution of multiple memory systems. *Psychological Review*, 94(4):439–454, 1987.
- [171] BF Skinner. A functional analysis of verbal behavior. *Psycholinguistics: a book of readings*, page 67, 1961.
- [172] S.A. Sloman. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3–22, 1996.
- [173] E.R. Smith and J. DeCoster. Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, 4(2):108, 2000.
- [174] P. Smolensky. On the proper treatment of connectionism. *Connectionism: Debates on psychological explanation*, 2:28–89, 1995.
- [175] MJ Spivey. *The continuity of mind*. Oxford University Press, USA, 2007.
- [176] M.J. Spivey, R. Dale, G. Knoblich, and M. Grosjean. Do curved reaching movements emerge from competing perceptions? a reply to van der wel et al. (2009). 2010.
- [177] M.J. Spivey, M. Grosjean, and G. Knoblich. Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10393, 2005.
- [178] M.J. Spivey and M.K. Tanenhaus. Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24:1521–1543, 1998.
- [179] D. Stanley, E. Phelps, and M. Banaji. The neural basis of implicit attitudes. *Current Directions in Psychological Science*, 17(2):164, 2008.

- [180] K.E. Stanovich and R.F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences*, 23(05):645–665, 2001.
- [181] K.E. Stephan, C.C. Hilgetag, GA Burns, M.A. O’Neill, M.P. Young, and R. K “otter. Computational analysis of functional connectivity between areas of primate cerebral cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355(1393):111, 2000.
- [182] F. Strack and R. Deutsch. Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8:220–247, 2004.
- [183] S.H. Strogatz. Non Linear Dynamics. *Adison Wesley*, 1994.
- [184] R.E. Suri and W. Schultz. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, 121(3):350–354, 1998.
- [185] R.E. Suri and W. Schultz. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91:871–890, 1999.
- [186] R.S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.
- [187] M.K. Tanenhaus, M.J. Spivey-Knowlton, K.M. Eberhard, and J.C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632, 1995.
- [188] D.R. Taylor and L.W. Aarssen. Complex competitive relationships among genotypes of three perennial grasses: implications for species coexistence. *American Naturalist*, pages 305–327, 1990.
- [189] G. Tesauro. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219, 1994.

- [190] G. Tononi, G.M. Edelman, and O. Sporns. Complexity and coherency: integrating information in the brain. *Trends in cognitive sciences*, 2(12):474–484, 1998.
- [191] M. Usher and J.L. McClelland. The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3):550–592, 2001.
- [192] R.R. Vallacher, A. Nowak, M. Froehlich, and M. Rockloff. The dynamics of self-evaluation. *Personality and Social Psychology Review*, 6(4):370–379, 2002.
- [193] R.R. Vallacher and A.E. Nowak. *Dynamical systems in social psychology*. Academic Press, 1994.
- [194] R.P.R.D. Van Der Wel, J.R. Eder, A.D. Mitchel, M.M. Walsh, and D.A. Rosenbaum. Trajectories emerging from discrete versus continuous processing models in phonological competitor tasks: A commentary on spivey, grosjean, and knoblich (2005). 2009.
- [195] G.C. Van Orden, J.G. Holden, and M.T. Turvey. Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132(3):331–350, 2003.
- [196] A. Vladimírsky. Static PDEs for time-dependent control problems. *Interfaces and Free Boundaries*, 8(3):281, 2006.
- [197] W. Von Hippel, L. Brener, and C. Von Hippel. Implicit prejudice toward injecting drug users predicts intentions to change jobs among drug and alcohol nurses. *Psychological Science*, 19(1):7, 2008.
- [198] C.J.C.H. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.
- [199] T.J. Wills, C. Lever, F. Cacucci, N. Burgess, and J. O’Keefe. Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873, 2005.

- [200] T.D. Wilson, S. Lindsey, and T.Y. Schooler. A model of dual attitudes. *Psychological review*, 107(1):101–126, 2000.
- [201] B. Wittenbrink, C.M. Judd, and B. Park. Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, 72(2):262–274, 1997.
- [202] B. Wittenbrink and N. Schwarz. *Implicit measures of attitudes*. The Guilford Press, 2007.
- [203] M.T. Wojnowicz, M.J. Ferguson, R. Dale, and M.J. Spivey. The self-organization of explicit attitudes. *Psychological Science*, 20(11):1428, 2009.
- [204] J. Yamanishi, M. Kawato, and R. Suzuki. Two coupled oscillators as a model for the coordinated finger tapping by both hands. *Biological Cybernetics*, 37(4):219–225, 1980.
- [205] M.P. Young. The organization of neural systems in the primate cerebral cortex. *Proceedings: Biological Sciences*, 252(1333):13–18, 1993.
- [206] R.B. Zajonc. Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, 10(6):224, 2001.
- [207] R.S. Zemel, P. Dayan, and A. Pouget. Probabilistic interpretation of population codes. *Neural Computation*, 10(2):403–430, 1998.